Sampled forms of functional PCA in reproducing kernel Hilbert spaces

Arash A. Amini[⋆] Martin J. Wainwright[†,⋆]

Department of Statistics[†], and
Department of Electrical Engineering and Computer Sciences[⋆]
UC Berkeley, Berkeley, CA 94720

September 16, 2011

Abstract

We consider the sampling problem for functional PCA (fPCA), where the simplest example is the case of taking time samples of the underlying functional components. More generally, model the sampling operation as a continuous linear map from $\mathcal{H}$ to $\mathbb{R}^m$, where the functional components to lie in some Hilbert subspace $\mathcal{H}$ of L^2, for example an RKHS of smooth functions. This model includes time and frequency sampling as special cases. In contrast to classical approach in fPCA in which access to entire functions is assumed, having a limited number m of functional samples places limitations on the performance of statistical procedures. We study these effects by analyzing the rate of convergence of an M-estimator for the subspace spanned by the leading components in a multi-spiked covariance model. The estimator takes the form of regularized PCA, and hence is computationally attractive. We analyze the behavior of this estimator within a non-asymptotic framework, and provide bounds that hold with high probability as a function the number of statistical samples n and the number of functional samples m. We also derive lower bounds showing that the rates obtained are minimax optimal.

1 Introduction

The statistical analysis of functional data, commonly known as functional data analysis (FDA), is an established area of statistics with a great number of practical applications (e.g., see the books [27, 28] and references therein for various examples). When the data is available as finely sampled curves, say in time, it is common to treat it as a collection of continuous-time curves or functions, each being observed in totality. These datasets are then termed "functional", and various statistical procedures applicable in finite dimensions can be extended to this functional setting. Among such procedures is principal component analysis (PCA), which is the focus of present work.

If one thinks of continuity as a mathematical abstraction of reality, then treating functional data as continuous curves is arguably a valid modeling device. However, in practice, one is faced with finite computational resources and is forced to implement a (finite-dimensional) approximation of true functional procedures by some sort of truncation of functions, for instance in the frequency domain. It is then important to understand the effects of this truncation on the statistical performance of the procedure. In other situations, for example in longitudinal data analysis [12], a continuous curve model is justified as a hidden underlying generating process to which one has access only through sparsely sampled, corrupted by noise perhaps, measurements in time. Studying how the time-sampling affects the estimation of the underlying functions in the presence of noise has some elements in common with that of the frequency-domain problem mentioned above.

The aim of this paper is to study effects of "sampling"—in a fairly general sense—on functional principal component analysis in smooth function spaces. We take a functional-theoretic approach to sampling by treating the sampling procedure as a (continuous) linear operator. This provides us with a notion of sampling general enough to treat both the frequency-truncation and time-sampling within a unified framework. We take as our smooth function space a Hilbert subspace $\mathcal{H}$ of $L^2[0,1]$ and denote the sampling operator by $\Phi : \mathcal{H} \to \mathbb{R}^m$. We assume that there are functions $x_i(t)$, $t \in [0,1]$ in $\mathcal{H}$ for $i = 1, \ldots, n$, generated i.i.d. from a probabilistic model (to be discussed). We then observe the collection $\{\Phi x_i\}_{i=1}^n \subset \mathbb{R}^m$ in noise. We refer to the index n as the number of *statistical samples*, and to the index m as the number of *functional samples*.

We analyze a natural M-estimator which takes the form of a regularized PCA in $\mathbb{R}^m$ and provide rates of convergence in terms of n and m. The eigen-decay of two operators govern the rates, the product of Φ and its adjoint and the product of the map embedding $\mathcal{H}$ in L^2 and its adjoint. Our focus will be on the setting where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), in which case the two eigen-decays are intimately related through the kernel function $(s,t) \mapsto \mathbb{K}(s,t)$. In such cases, the two components of the rate interact and give rise to optimal values for the number of functional samples (m) in terms of the number of statistical samples (n) or vice versa. This has practical appeal in cases where obtaining either type of samples is costly.

Our model for the functions $\{x_i\}$ will be an extension to function spaces of the *spiked covariance model* introduced by Johnstone and his collaborators [17, 18], and studied by various authors (e.g., [18, 24, 1]). We consider such models with r components, each lying within the Hilbert ball $\mathbb{B}_{\mathcal{H}}(\rho)$ of radius ρ, with the goal of recovering the r-dimensional subspace spanned by the spiked components in this functional model. We analyze our M-estimators within a high-dimensional framework that allows both the number of statistical samples n and the number of functional samples m to diverge together. Our main theoretical contributions are to derive non-asymptotic bounds on the estimation error as a function of the pair (m,n), which are shown to be sharp (minimax-optimal). Although our rates also explicitly track the number of components r and the smoothness parameter ρ, we do not make any effort to obtain optimal dependence on these parameters.

The general asymptotic properties of PCA in function spaces have been investigated by various authors (e.g., [10, 7, 15].) Accounting for smoothness of functions by introducing various roughness/smoothness penalties is a standard approach, used in the papers [29, 25, 30, 6] among others. The problem of principal component analysis for sampled functions, with a similar functional-theoretic perspective, is discussed by Besse and Ramsey [4] for the noiseless case. A more recent line of work is devoted to the case of functional PCA with noisy sampled functions [8, 34, 16]. Cardot [8] considers estimation via spline-based approximation, and derives MISE rates in terms of various parameters of the model. Hall et al. [16] study estimation via local linear smoothing, and establish minimax-optimality in certain settings that involve a fixed number of functional samples. Both papers [8, 16] demonstrate trade-offs between the numbers of statistical and functional samples; we refer the reader to Hall et al. [16] for an illuminating discussion of connections between FDA and LDA approaches (i.e. having full versus sampled functions), which inspired much of the present work. We note that the regularization present in our M-estimator is closely related to classical roughness penalties [29, 30] in the special case of spline kernels, although the discussion there applies to fully-observed functions, as opposed to the sampled models considered here.

As mentioned above, our sampled model resembles very much that of spiked covariance model for high-dimensional principal component analysis. A line of work on this model has treated various types of sparsity conditions on the eigenfunctions [18, 24, 1]; in contrast, here the smoothness condition on functional components translates into an ellipsoid condition on the vector principal components. Perhaps an even more significant difference is that, here, the effective scaling of noise in $\mathbb{R}^m$ is substantially smaller in some cases (e.g., the case of time sampling). This could explain why the difficulty of "high-dimensional" setting is not observed in such cases as one lets $m, n \to \infty$. On the other hand, a difficulty particular to our sampled model is the lack of orthonormality between components (after sampling) which leads to identifiability issues; it also makes recovering individual components difficult. In order to derive non-asymptotic bounds on our M-estimator, we exploit various techniques from empirical process theory (e.g., [31]), as well as the concentration of measure (e.g., [20]). We also exploit recent work [23] on the localized Rademacher complexities of unit balls in a reproducing kernel Hilbert space, as well as techniques from non-asymptotic random matrix theory, as discussed in Davidson and Szarek [11], in order to control various norms of random marices. These techniques allow us to obtain finite-sample bounds that hold with high probability, and are specified explicitly in terms of the pair (m, n), and the underlying smoothness of the Hilbert space.

The remainder of this paper is organized as follows. Section 2 is devoted to background material on reproducing kernel Hilbert spaces, adjoints of operators, as well as the class of sampled functional models that we study in this paper. In Section 3, we describe M-estimators for sampled functional PCA, and discuss various implementation details. Section 4 is devoted to the statements of our main results, and discussion of their consequences for particular sampling models. In subsequent sections, we provide the proofs of our results, with some more technical aspects deferred to the appendices. Section 5 is devoted to bounds on the subspace-based error, whereas Section 6 is devoted to bounds on error in the function space. Section 7 provides matching lower bounds on the minimax error, showing that our analysis is sharp. We conclude with a discussion in Section 8.

Notation. We will use $|\!|\!| \cdot |\!|\!|_{HS}$ to denote the Hilbert-Schmidt norm of an operator or a matrix. The corresponding inner product is denoted as $\langle\!\langle \cdot, \, \cdot \rangle\!\rangle$. If T is an operator on a Hilbert space $\mathcal{H}$ with an orthonormal basis $\{e_j\}$, then $|\!|\!| T |\!|\!|_{HS}^2 = \sum_j \|Te_j\|_{\mathcal{H}}^2$. For a matrix $A = (a_{ij})$, we have $|\!|\!| A |\!|\!|_{HS}^2 = \sum_{i,j} |a_{ij}|^2$.

2 Background and problem set-up

In this section, we begin by introducing background on reproducing kernel Hilbert spaces, as well as linear operators and their adjoints. We then introduce the functional and observation model that we study in this paper, and conclude with discussion of some approximation-theoretic issues that play an important role in parts of our analysis.

2.1 Reproducing Kernel Hilbert Spaces

We begin with a quick overview of some standard properties of reproducing kernel Hilbert spaces; we refer the reader to the books [32, 14] for more details. A reproducing kernel Hilbert space (or RKHS for short) is a Hilbert space $\mathcal{H}$ of functions $f : T \to \mathbb{R}$ that is equipped with an associated kernel $\mathbb{K} : T \times T \to \mathbb{R}$. We assume the kernel to be continuous, and the set $T \subset \mathbb{R}^d$ to be compact. For concreteness, we think of $T = [0, 1]$ throughout this paper, but any compact set of $\mathbb{R}^d$ suffices. For each $t \in T$, the function $R_t := \mathbb{K}(\cdot, t)$ belongs to the Hilbert space $\mathcal{H}$, and it acts as the *representer of evaluation*, meaning that $\langle f, R_t \rangle_{\mathcal{H}} = f(t)$ for all $f \in \mathcal{H}$.

The kernel $\mathbb{K}$ defines an integral operator $\mathcal{T}_{\mathbb{K}}$ on $L^2(T)$, mapping the function f to the function $g(s) = \int_T K(s,t) f(t) dt$. By the spectral theorem in Hilbert spaces, this operator can be associated with a sequence of eigenfunctions $\psi_k, k = 1, 2, \ldots$ in $\mathcal{H}$, orthogonal in $\mathcal{H}$ and orthonormal in $L^2(T)$, and a sequence of non-negative eigenvalues $\mu_1 \geq \mu_2 \geq \cdots$. Most useful for this paper is the fact that any function $f \in \mathcal{H}$ has an expansion in terms of these eigenfunctions and eigenvalues, namely

$$f = \sum_{k=1}^{\infty} \sqrt{\mu_k} \alpha_k \psi_k \tag{1}$$

for some $(\alpha_k) \in \ell^2$. In terms of this expansion, we have the representations $\|f\|_{\mathcal{H}}^2 = \sum_{k=1}^{\infty} \alpha_k^2$ and $\|f\|_{L^2}^2 = \sum_{k=1}^{\infty} \mu_k \alpha_k^2$. Many of our results involve the decay rate of these eigenvalues: in particular, for some parameter $\alpha > 1/2$, we say that the kernel operator has eigenvalues with *polynomial-α decay* if there is a constant $c > 0$ such that

$$\mu_k \leq \frac{c}{k^{2\alpha}} \quad \text{for all } k = 1, 2, \ldots. \tag{2}$$

Let us consider an example to illustrate.

Example 1 (Sobolev class with smoothness $\alpha = 1$)**.** In the case $T = [0, 1]$ and $\alpha = 1$, we can consider the kernel function $\mathbb{K}(s,t) = \min\{s,t\}$. As discussed in Appendix A, this kernel generates the class of functions

$$\mathcal{H} := \big\{ f \in L^2([0,1]) \;\mid\; f(0) = 0,\ f \text{ absolutely continuous and } f' \in L^2([0,1]) \big\}.$$

The class $\mathcal{H}$ is an RKHS with inner product $\langle f, g \rangle_{\mathcal{H}} = \int_0^1 f'(t) g'(t) dt$, and the ball $\mathbb{B}_{\mathcal{H}}(\rho)$ corresponds to a Sobolev space with smoothness $\alpha = 1$. The eigen-decomposition of the kernel integral operator is

$$\mu_k = \Big[\frac{(2k-1)\pi}{2} \Big]^{-2}, \quad \psi_k(t) = \sqrt{2} \sin\big(\mu_k^{-1/2} t\big), \quad k = 1, 2, \ldots. \tag{3}$$

Consequently, this class has polynomial decay with parameter $\alpha = 1$.

We note that there are natural generalizations of this example to $\alpha = 2, 3, \ldots$, corresponding to the Sobolev classes of α-times differentiable functions (e.g., see the books [14, 3]).

In this paper, the operation of generalized sampling is defined in terms of a bounded linear operator $\Phi : \mathcal{H} \to \mathbb{R}^m$ on the Hilbert space. Its adjoint is a mapping $\Phi^* : \mathbb{R}^m \to \mathcal{H}$, defined by the relation $\langle \Phi f, a \rangle_{\mathbb{R}^m} = \langle f, \Phi^* a \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $a \in \mathbb{R}^m$. In order to compute a representation of the adjoint, we note that by the Riesz representation theorem, the j-th coordinate of this mapping—namely, $f \mapsto [\Phi f]_j$—can be represented as an inner product $\langle \phi_j, f \rangle_{\mathcal{H}}$, for some element $\phi_j \in \mathcal{H}$, and we can write

$$\Phi f = \begin{bmatrix} \langle \phi_1, f \rangle_{\mathcal{H}} & \langle \phi_2, f \rangle_{\mathcal{H}} & \cdots & \langle \phi_m, f \rangle_{\mathcal{H}} \end{bmatrix}^T. \tag{4}$$

Consequently, we have $\langle \Phi(f), a \rangle_{\mathbb{R}^m} = \sum_{j=1}^m a_j \langle \phi_j, f \rangle_{\mathcal{H}} = \langle \sum_{j=1}^m a_j \phi_j, f \rangle_{\mathcal{H}}$, so that for any $a \in \mathbb{R}^m$, the adjoint can be written as

$$\Phi^* a = \sum_{j=1}^m a_j \phi_j. \tag{5}$$

This adjoint operator plays an important role in our analysis.

2.2 Functional model and observations

Let $s_1 \geq s_2 \geq s_3 \geq \cdots \geq s_r > 0$ be a fixed sequence of positive numbers, and let $\{f_j^*\}_{j=1}^r$ be a fixed sequence of functions orthonormal in $L^2[0,1]$. Consider a collection of n i.i.d. random functions $\{x_1, \ldots, x_n\}$, generated according to the model

$$x_i(t) = \sum_{j=1}^{r} s_j \beta_{ij} f_j^*(t), \quad \text{for } i = 1, \ldots, n, \tag{6}$$

where $\{\beta_{ij}\}$ are i.i.d. $N(0,1)$ across all pairs (i,j). This model corresponds to a finite-rank instantiation of functional PCA, in which the goal is to estimate the span of the unknown eigenfunctions $\{f_j^*\}_{j=1}^r$. Typically, these eigenfunctions are assumed to satisfy certain smoothness conditions; in this paper, we model such conditions by assuming that the eigenfunctions belong to a reproducing kernel Hilbert space $\mathcal{H}$ embedded within $L^2[0,1]$; more specifically, they lie in some ball in $\mathcal{H}$,

$$\|f_j^*\|_{\mathcal{H}} \leq \rho, \quad j = 1, \ldots, r. \tag{7}$$

For statistical problems involving estimation of functions, the random functions might only be observed at certain times $(t_1, \ldots, t_m)$, such as in longitudinal data analysis, or we might collect only projections of each x_i in certain directions, such as in tomographic reconstruction. More concretely, in a *time-sampling model*, we observe m-dimensional vectors of the form

$$y_i = \begin{bmatrix} x_i(t_1) & x_i(t_2) & \cdots & x_i(t_m) \end{bmatrix}^T + \sigma_0 w_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{8}$$

where $\{t_1, t_2, \ldots, t_m\}$ is a fixed collection of design points, and $w_i \in \mathbb{R}^m$ is a noise vector. Another observation model is the *basis truncation model* in which we observe the projections of f onto the first m basis functions $\{\psi_j\}_{j=1}^m$ of the kernel operator—namely,

$$y_i = \begin{bmatrix} \langle \psi_1, x_i \rangle_{L^2} & \langle \psi_2, x_i \rangle_{L^2} & \cdots & \langle \psi_m, x_i \rangle_{L^2} \end{bmatrix}^T + \sigma_0 w_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{9}$$

where $\langle \cdot, \cdot \rangle_{L^2}$ represents the inner product in $L^2[0,1]$.

In order to model these and other scenarios in a unified manner, we introduce a linear operator Φ_m that maps any function x in the Hilbert space to a vector $\Phi_m(x)$ of m samples, and then consider the linear observation model

$$y_i = \Phi_m(x_i) + \sigma_m w_i, \qquad \text{for } i = 1, 2, \ldots, n. \tag{10}$$

This model (10) can be viewed as a functional analog of the spiked covariance models introduced by Johnstone [17, 18] as an analytically-convenient model for studying high-dimensional effects in classical PCA.

Both the time-sampling (8) and frequency truncation (9) models can be represented in this way, for appropriate choices of the operator Φ_m. Recall the representation (4) of Φ_m in terms of the functions $\{\phi_j\}_{j=1}^m$.

- For the time sampling model (8), we set $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$, so that by the reproducing property of the kernel, we have $\langle \phi_j, f \rangle_{\mathcal{H}} = f(t_j)/\sqrt{m}$ for all $f \in \mathcal{H}$, and $j = 1, 2, \ldots m$. With these choices, the operator Φ_m maps each $f \in \mathcal{H}$ to the m-vector of rescaled samples $\frac{1}{\sqrt{m}} \begin{bmatrix} f(t_1) & \cdots & f(t_m) \end{bmatrix}^T$. Defining the rescaled noise $\sigma_m = \frac{\sigma_0}{\sqrt{m}}$ yields an instantiation of the model (10) which is equivalent to time-sampling (8).

- For the basis truncation model (9), we set $\phi_j = \mu_j \psi_j$ so that the operator Φ maps each function $f \in \mathcal{H}$ to the vector of basis coefficients$\big[\langle \psi_1, f\rangle_{L^2} \quad \cdots \quad \langle \psi_m, f\rangle_{L^2}\big]^T$. Setting $\sigma_m = \sigma_0$ then yields another instantiation of the model (10), this one equivalent to basis truncation (9).

A remark on notation before proceeding: in the remainder of the paper, we use (Φ, σ) as shorthand notation for (Φ_m, σ_m), since the index m should be implicitly understood throughout our analysis.

In this paper, we provide and analyze estimators for the r-dimensional eigen-subspace spanned by $\{f_j^*\}$, in both the sampled domain $\mathbb{R}^m$, and in the functional domain. To be more specific, for $j = 1, \dots, r$, define the vectors $z_j^* := \Phi f_j^* \in \mathbb{R}^m$, and the subspaces

$$\mathfrak{Z}^* := \operatorname{span}\{z_1^*, \dots, z_r^*\} \subset \mathbb{R}^m, \quad \text{and} \quad \mathfrak{F}^* := \operatorname{span}\{f_1^*, \dots, f_r^*\} \subset \mathcal{H},$$

and let $\widehat{\mathfrak{Z}}$ and $\widehat{\mathfrak{F}}$ denote the corresponding estimators. In order to measure the performance of the estimators, we will use projection-based distances between subspaces. In particular, let $P_{\mathfrak{Z}^*}$ and $P_{\widehat{\mathfrak{Z}}}$ be orthogonal projection operators into $\mathfrak{Z}^*$ and $\widehat{\mathfrak{Z}}$, respectively, considered as subspaces of $\ell_2^m := (\mathbb{R}^m, \|\cdot\|_2)$. Similarly, let $P_{\mathfrak{F}^*}$ and $P_{\widehat{\mathfrak{F}}}$ be orthogonal projection operators into $\mathfrak{F}^*$ and $\widehat{\mathfrak{F}}$, respectively, considered as subspaces of $(\mathcal{H}, \|\cdot\|_{L^2})$. We are interested in bounding the deviations

$$\mathrm{d}_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) := |\!|\!| P_{\widehat{\mathfrak{Z}}} - P_{\mathfrak{Z}^*} |\!|\!|_{HS}, \quad \text{and} \quad \mathrm{d}_{HS}(\widehat{\mathfrak{F}}, \mathfrak{F}^*) := |\!|\!| P_{\widehat{\mathfrak{F}}} - P_{\mathfrak{F}^*} |\!|\!|_{HS}, \tag{11}$$

where $|\!|\!| \cdot |\!|\!|_{HS}$ is the Hilbert-Schmidt norm of an operator (or matrix).

2.3 Approximation-theoretic quantities

One object that plays an important role in our analysis is the matrix $K := \Phi\Phi^* \in \mathbb{R}^{m \times m}$. From the form of the adjoint, it can be seen that $[K]_{ij} = \langle \phi_i, \phi_j \rangle_{\mathcal{H}}$. For future reference, let us compute this matrix for the two special cases of linear operators considered thus far.

- For the time sampling model (8), we have $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$ for all $j = 1, \dots, m$, and hence $[K]_{ij} = \frac{1}{m}\langle \mathbb{K}(\cdot, t_i), \mathbb{K}(\cdot, t_j)\rangle_{\mathcal{H}} = \frac{1}{m}\mathbb{K}(t_i, t_j)$, using the reproducing property of the kernel.
- For the basis truncation model (9), we have $\phi_j = \mu_j \psi_j$, and hence $[K]_{ij} = \langle \mu_i \psi_i, \mu_j \psi_j \rangle_{\mathcal{H}} = \mu_i \delta_{ij}$. Thus, in this special case, we have $K = \operatorname{diag}(\mu_1, \dots, \mu_m)$.

In general, the matrix K is a type of Gram matrix, and so is symmetric and positive semidefinite. We assume throughout this paper that the functions $\{\phi_j\}_{j=1}^m$ are linearly independent in $\mathcal{H}$, which implies that K is strictly positive definite. Consequently, it has a set of eigenvalues which can be ordered as

$$\widehat{\mu}_1 \;\geq\; \widehat{\mu}_2 \;\geq\; \dots \;\geq \widehat{\mu}_m > 0. \tag{12}$$

Under this condition, we may use K to define a norm on $\mathbb{R}^m$ via $\|z\|_K^2 := z^T K^{-1} z$. Moreover, we have the following interpolation lemma, which is proved Appendix B.1:

Lemma 1. *For any $f \in \mathcal{H}$, we have $\|\Phi f\|_K \;\leq\; \|f\|_{\mathcal{H}}$, with equality if and only if $f \in \operatorname{Ra}(\Phi^*)$. Moreover, for any $z \in \mathbb{R}^m$, the function $g = \Phi^* K^{-1} z$ has smallest Hilbert norm of all functions satisfying $\Phi g = z$, and is the unique function with this property.*

This lemma is useful in constructing a function-based estimator, as will be clarified in Section 3.

In our analysis of the functional error $\mathrm{d}_{HS}(\widehat{\mathfrak{F}}, \mathfrak{F}^*)$, a number of approximation-theoretic quantities play an important role. As a mapping from an infinite-dimensional space $\mathcal{H}$ to $\mathbb{R}^m$, the operator Φ has a non-trivial nullspace. Given the observation model (10), we receive no information about any component of a function f^* that lies within this nullspace. For this reason, we define the width of the nullspace in the L^2-norm, namely the quantity

$$N_m(\Phi) := \sup\big\{\|f\|_{L^2}^2 \;\mid\; f \in \mathrm{Ker}(\Phi), \|f\|_{\mathcal{H}} \leq 1\big\}. \tag{13}$$

In addition, the observation operator Φ induces a semi-norm on the space $\mathcal{H}$, defined by

$$\|f\|_{\Phi}^2 := \|\Phi f\|_2^2 \;=\; \sum_{j=1}^m [\Phi f]_j^2. \tag{14}$$

It is of interest to assess how well this semi-norm approximates the L^2-norm. Accordingly, we define the quantity

$$D_m(\Phi) := \sup_{\substack{f \in \mathrm{Ra}(\Phi^*) \\ \|f\|_{\mathcal{H}} \leq 1}} \big|\|f\|_{\Phi}^2 - \|f\|_{L^2}^2\big|, \tag{15}$$

which measures the worst-case gap between these two (semi)-norms, uniformly over the Hilbert ball of radius one, restricted to the subspace of interest $\mathrm{Ra}(\Phi^*)$. Given knowledge of the linear operator Φ, the quantity $D_m(\Phi)$ can be computed in a relatively straightforward manner. In particular, recall the definition of the matrix K, and let us define a second matrix $\Theta \in \mathbb{S}_+^m$ with entries $\Theta_{ij} := \langle \varphi_i, \varphi_j \rangle_{L^2}$.

Lemma 2. *We have the equivalence*

$$D_m(\Phi) = |\!|\!| K - K^{-1/2}\Theta K^{-1/2} |\!|\!|_2, \tag{16}$$

where $|\!|\!| \cdot |\!|\!|_2$ *denotes the* ℓ_2*-operator norm.*

See Appendix B.2 for the proof of this claim.

3 M-estimator and implementation

With this background in place, we now turn to the description of our M-estimator, as well as practical details associated with its implementation.

3.1 M-estimator

We begin with some preliminaries on notation, and our representation of subspaces. For each $j = 1, \ldots, m$, define the vector $z_j^* := \Phi f_j^*$, corresponding to the image of the function f_j^* under the observation operator. We let $\mathfrak{Z}^*$ denote the r-dimensional subspace of $\mathbb{R}^m$ spanned by $\{z_1^*, \ldots, z_r^*\}$, where $z_j^* = \Phi f_j^*$. Our initial goal is to construct an estimate $\widehat{\mathfrak{Z}}$, itself an r-dimensional subspace, of the unknown subspace $\mathfrak{Z}^*$.

We represent subspaces by elements of the Stiefel manifold $V_r(\mathbb{R}^m)$, which consists of of $m \times r$ matrices Z with orthonormal columns

$$V_r(\mathbb{R}^m) := \big\{Z \in \mathbb{R}^{m \times r} \;\mid\; Z^T Z = I_r\big\}.$$

A given matrix Z acts as a representative of the subspace spanned by its columns, denoted by $\mathrm{col}(Z)$. For any $U \in V_r(\mathbb{R}^r)$, the matrix ZU also belongs to the Stiefel manifold, and since $\mathrm{col}(Z) = \mathrm{col}(ZU)$, we may call ZU a version of Z. We let $P_Z = ZZ^T \in \mathbb{R}^{m \times m}$ be the orthogonal projection onto $\mathrm{col}(Z)$. For two matrices $Z_1, Z_2 \in V_r(\mathbb{R}^m)$, we measure the distance between the associated subspaces via $\mathrm{d}_{HS}(Z_1, Z_2) := |\!|\!| P_{Z_1} - P_{Z_2} |\!|\!|_{HS}$, where $|\!|\!| \cdot |\!|\!|_{HS}$ is the Hilbert-Schmidt (or Frobenius) matrix norm.

3.1.1 Subspace-based estimator

With this notation, we now specify an M-estimator for the subspace $\mathfrak{Z}^* = \mathrm{span}\{z_1^*, \ldots, z_r^*\}$. Let us begin with some intuition. Given the n samples $\{y_1, \ldots, y_n\}$, let us define the $m \times m$ sample covariance matrix $\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n y_i y_i^T$. Given the observation model (10), a straightforward computation shows that

$$\mathbb{E}[\widehat{\Sigma}_n] = \sum_{j=1}^r s_j^2 z_j^* (z_j^*)^T + \sigma_m^2 I_m. \tag{17}$$

Thus, as n becomes large, we expect that the top r eigenvectors of $\widehat{\Sigma}_n$ might give a good approximation to $\mathrm{span}\{z_1^*, \ldots, z_r^*\}$. By the Courant-Fischer variational representation, these r eigenvectors can be obtained by maximizing the objective function

$$\langle\!\langle \widehat{\Sigma}_n,\; P_Z \rangle\!\rangle := \mathrm{tr}(\widehat{\Sigma}_n Z Z^T)$$

over all matrices $Z \in V_r(\mathbb{R}^m)$.

However, this approach fails to take into account the smoothness constraints that the vectors $z_j^* = \Phi f_j^*$ inherit from the smoothness of the eigenfunctions f_j^*. Since $\|f_j^*\|_{\mathcal{H}} \leq \rho$ by assumption, Lemma 1 implies that

$$\|z_j^*\|_K^2 = (z_j^*)^T K^{-1} z_j^* \;\leq\; \|f_j^*\|_{\mathcal{H}}^2 \;\leq\; \rho^2 \quad \text{for all } j = 1, 2, \ldots, r.$$

Consequently, if we define the matrix $Z^* := \begin{bmatrix} z_1^* & \cdots & z_r^* \end{bmatrix} \in \mathbb{R}^{m \times r}$, then it must satisfy the *trace smoothness condition*

$$\langle\!\langle K^{-1},\; Z^*(Z^*)^T \rangle\!\rangle = \sum_{j=1}^r (z_j^*)^T K^{-1} z_j^* \;\leq\; r\rho^2. \tag{18}$$

This calculation motivates the constraint $\langle\!\langle K^{-1},\; P_Z \rangle\!\rangle \leq 2r\rho^2$ in our estimation procedure.

Based on the preceding intuition, we are led to consider the optimization problem

$$\widehat{Z} \in \arg \max_{Z \in V_r(\mathbb{R}^m)} \Big\{ \langle\!\langle \widehat{\Sigma}_n,\; P_Z \rangle\!\rangle \;\mid\; \langle\!\langle K^{-1},\; P_Z \rangle\!\rangle \leq 2r\rho^2 \Big\}, \tag{19}$$

where we recall that $P_Z = ZZ^T \in \mathbb{R}^{m \times m}$. Given any optimal solution $\widehat{Z}$, we return the subspace $\widehat{\mathfrak{Z}} = \mathrm{col}(\widehat{Z})$ as our estimate of $\mathfrak{Z}^*$. As discussed at more length in Section 3.2, it is straightforward to compute $\widehat{Z}$ in polynomial time. The reader might wonder why we have included an additional factor of two in this trace smoothness condition. This slack is actually needed due to the potential infeasibility of the matrix Z^* for the program (19), which arises since the columns Z^* are not guaranteed to be orthonormal. As shown by our analysis, the additional slack allows us to find a matrix $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$ that spans the same subspace as Z^*, and is also feasible for the program (19). More formally, we have:

Lemma 3. *Under condition (26b), there exists a matrix $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$ such that*

$$\mathrm{Ra}(\widetilde{Z}^*) = \mathrm{Ra}(Z^*), \quad \textit{and} \quad \langle\!\langle K^{-1},\ \widetilde{Z}^*(\widetilde{Z}^*)^T \rangle\!\rangle \le 2r\rho^2. \tag{20}$$

See Appendix B.3 for the proof of this claim.

3.1.2 The functional estimate $\widehat{\mathfrak{F}}$

Having obtained an estimate[1] $\widehat{\mathfrak{Z}} = \mathrm{span}\{\widehat{z}_1, \dots, \widehat{z}_r\}$ of $\mathfrak{Z}^* = \mathrm{span}\{z_1^*, \dots, z_r^*\}$, we now need to construct a r-dimensional subspace $\widehat{\mathfrak{F}}$ of the Hilbert space as an estimate of $\mathfrak{F}^* = \mathrm{span}\{f_1^*, \dots, f_r^*\}$. We do so using the interpolation suggested by Lemma 1. For each $j = 1, \dots, r$, define the function

$$\widehat{f}_j := \Phi^* K^{-1} \widehat{z}_j \ = \ \sum_{i=1}^{m} (K^{-1}\widehat{z}_j)_i \ \phi_i. \tag{21}$$

Since $K = \Phi\Phi^*$ by definition, this construction ensures that $\Phi\widehat{f}_j = \widehat{z}_j$. Moreover, Lemma 1 guarantees that $\widehat{f}_j$ has the minimal Hilbert norm (and hence is smoothest in a certain sense) over all functions that have this property. Finally, since Φ is assumed to be surjective (equivalently, K assumed invertible), $\Phi^* K^{-1}$ maps linearly independent vectors to linearly independent functions, and hence preserves dimension. Consequently, the space $\widehat{\mathfrak{F}} := \mathrm{span}\{\widehat{f}_1, \dots, \widehat{f}_r\}$ is an r-dimensional subspace of $\mathcal{H}$ which we take as our estimate of $\mathfrak{F}^*$.

3.2 Implementation details

In this section, we consider some practical aspects of implementing the M-estimator, and present some simulations to illustrate its qualitative properties. We begin by observing that once the subspace vectors $\{\widehat{z}_j\}_{j=1}^r$ have been computed, then it is straightforward to compute the function estimates $\{\widehat{f}_j\}_{j=1}^r$, as weighted combinations of the functions $\{\phi_j\}_{j=1}^m$. Accordingly, we focus our attention on solving the program (19).

On the surface, the problem (19) might appear non-convex, due to the Stiefel manifold constraint. However, it can be reformulated as a semidefinite program (SDP), a well-known class of convex programs, as clarified in the following:

Lemma 4. *The problem* (19) *is equivalent to solving the SDP*

$$\widehat{X} \in \arg\max_{X \succeq 0} \langle\!\langle \widehat{\Sigma}_n,\ X \rangle\!\rangle \quad \textit{such that } |\!|\!| X |\!|\!|_2 \le 1,\ \mathrm{tr}(X) = r,\ \textit{and } \langle\!\langle K^{-1},\ X \rangle\!\rangle \le 2r\rho^2, \tag{22}$$

[1] Here, $\{\widehat{z}_j\}_{j=1}^r \subset \mathbb{R}^m$ is any collection of vectors that span $\widehat{\mathfrak{Z}}$. As we are ultimately only interested in the resulting functional "subspace", it does not matter which particular collection we choose.

for which there always exists an optimal rank r solution. Moreover, by Lagrangian duality, for some $\beta > 0$, the problem is equivalent to

$$\widehat{X} \in \arg\max_{X \succeq 0} \langle\langle \widehat{\Sigma}_n - \beta K^{-1},\ X \rangle\rangle \quad \textit{such that } |||X|||_2 \leq 1 \textit{ and } \operatorname{tr}(X) = r, \tag{23}$$

which can be solved by an eigendecomposition of $\widehat{\Sigma}_n - \beta K^{-1}$.

As a consequence, for a given Lagrange multiplier β, the regularized form of the estimator can be solved with the cost of solving an eigenvalue problem. For a given constraint $2r\rho^2$, the appropriate value of β can be found by a path-tracing algorithm, or a simple dyadic splitting approach.

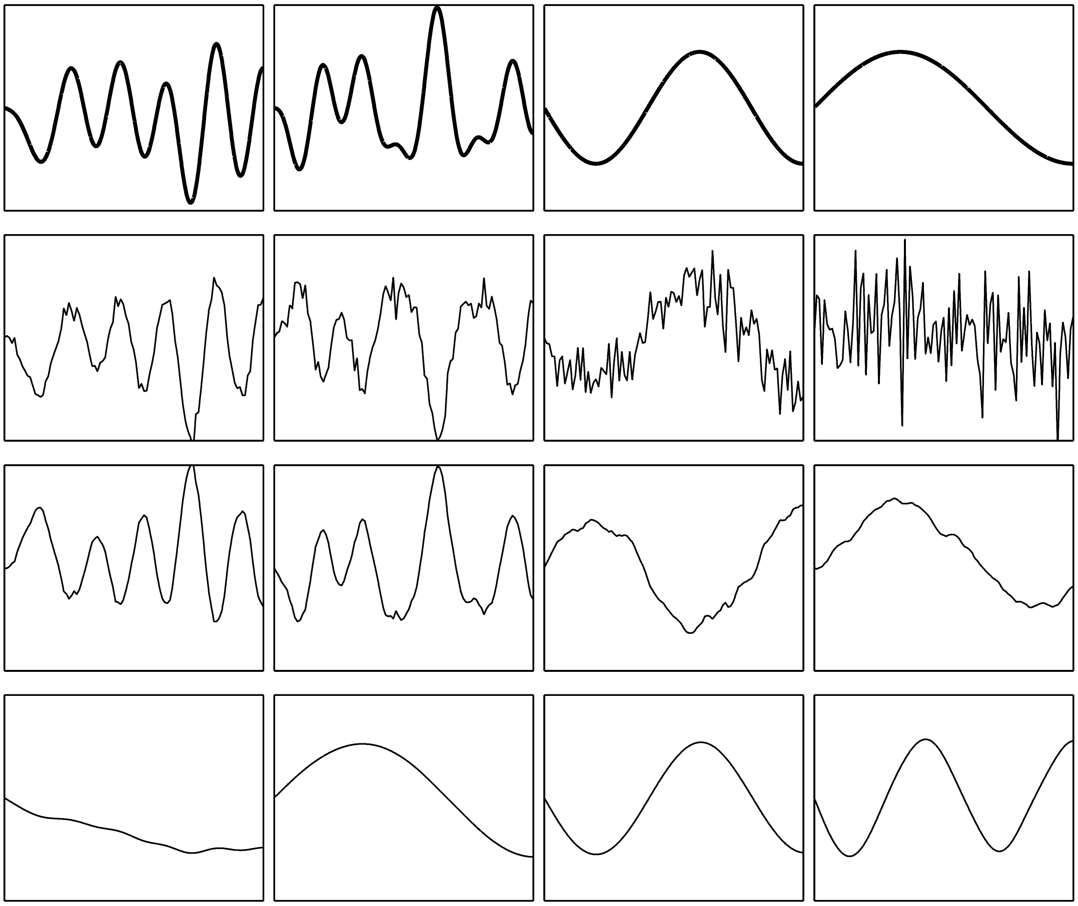

Figure 1: Regularized PCA for time sampling in first-order Sobolev RKHS. Top row shows, from left to right, plots of the $r = 4$ "true" principal components $f_1^*, \ldots, f_4^*$ with signal-to-noise ratios $s_1 = 1$, $s_2 = 0.5$, $s_3 = 0.25$ and $s_4 = 0.125$, respectively. The number of statistical and functional samples are $n = 75$ and $m = 100$. Subsequent rows show the corresponding estimators $\widehat{f}_1, \ldots, \widehat{f}_4$ obtained by applying the regularized form (23).

In order to illustrate the estimator, we consider the time sampling model (8), with uniformly spaced samples, in the context of a first-order Sobolev RKHS (with kernel function $\mathbb{K}(s,t) = \min(s,t)$). The parameters of the model are taken to be $r = 4$, $(s_1, s_2, s_3, s_4) =$

(1, 0.5, 0.25, 0.125), $\sigma_0 = 1$, $m = 100$ and $n = 75$. The regularized form (23) of the estimator is applied and the results are shown in Fig. 1. The top row corresponds to the four "true" signals $\{f_j^*\}$, the leftmost being f_1^* (i.e. having the highest signal-to-noise ratio.) and the rightmost f_4^*. The subsequent rows show the corresponding estimates $\{\widehat{f}_j\}$, obtained using different values of β. The second, third and fourth rows correspond to $\beta = 0$, $\beta = 0.0052$ and $\beta = 0.83$.

One observes that without regularization ($\beta = 0$), the estimates for two weakest signals (f_3^* and f_4^*) are poor. The case $\beta = 0.0052$ is roughly the one which achieves the minimum for the dual problem. One observes that the quality of the estimates of the signals, and in particular the weakest ones, are considerably improved. The optimal (oracle) value of β, that is the one which achieves the minimum error between $\{f_j^*\}$ and $\{\widehat{f}_j\}$, is $\beta = 0.0075$ in this problem. The corresponding estimates are qualitatively similar to those of $\beta = 0.0052$ and are not shown.

The case $\beta = 0.83$ shows the effect of over-regularization. It produces very smooth signals and although it fails to reveal f_1^* and f_2^*, it reveals highly accurate versions of f_3^* and f_4^*. It is also interesting to note that the smoothest signal, f_4^*, now occupies the position of the second (estimated) principal component. That is, the regularized PCA sees an effective signal-to-noise ratio which is influenced by smoothness. This suggests a rather practical appeal of the method in revealing smooth signals embedded in noise. One can vary β from zero upward and if some patterns seem to be present for a wide range of β (and getting smoother as β is increased), one might suspect that they are indeed present in data but masked by noise.

4 Main results

We now turn to the statistical analysis of our estimators, in particular deriving high-probability upper bounds on the error of the subspace-based estimate $\widehat{\mathfrak{Z}}$, and the functional estimate $\widehat{\mathfrak{F}}$. In both cases, we begin by stating general theorems that applies to arbitrary linear operators Φ—Theorems 1 and 2 respectively—and then derive a number of corollaries for particular instantiations of the observation operator.

4.1 Subspace-based estimation rates (for $\widehat{\mathfrak{Z}}$)

We begin by stating high-probability upper bounds on the error $\mathrm{d}_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*)$ of the subspace-based estimates. Our rates are stated in terms of a function that involves the eigenvalues of the matrix $K = \Phi\Phi^* \in \mathbb{R}^m$, ordered as $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots \geq \widehat{\mu}_m > 0$. Consider the function $\mathcal{F} : \mathbb{R}_+ \to \mathbb{R}_+$ given by

$$\mathcal{F}(t) := \Big[\sum_{j=1}^{m} \min\{t^2, r\rho^2 \widehat{\mu}_j\}\Big]^{1/2}. \tag{24}$$

As will be clarified in our proofs, this function provides a measure of the statistical complexity of the function class $\mathrm{Ra}(\Phi^*) = \{f \in \mathcal{H} \mid f = \sum_{j=1}^m a_j \phi_j \text{ for some } a \in \mathbb{R}^m\}$.

We require a few regularity assumptions. Define the quantity

$$C_m(f^*) := \max_{1 \leq i,j \leq r} \big|\langle f_i^*, f_j^*\rangle_\Phi - \delta_{ij}\big| \;=\; \max_{1 \leq i,j \leq r} \big|\langle z_i^*, z_j^*\rangle_{\mathbb{R}^m} - \delta_{ij}\big|, \tag{25}$$

which measures the departure from orthonormality of the vectors $z_j^* := \Phi f_j^*$ in $\mathbb{R}^m$. A straightforward argument using a polarization identity shows that $C_m(f^*)$ is upper bounded (up to a constant factor) by the uniform quantity $D_m(\Phi)$, as defined in equation (15). Recall that the random functions are generated according to the model $x_i = \sum_{j=1}^r s_j \beta_{ij} f_j^*$, where the signal strengths are ordered as $1 = s_1 \geq s_2 \geq \cdots \geq s_r > 0$, and that σ_m denotes the noise standard deviation in the observation model (10).

In terms of these quantities, we require the following assumptions:

$$\textbf{(A1)} \quad \frac{s_r^2}{s_1^2} \;\geq\; \frac{1}{2}, \quad \text{and} \quad \sigma_0^2 := \sup_m \sigma_m^2 \;\leq\; \kappa s_1^2, \tag{26a}$$

$$\textbf{(A2)} \quad C_m(f^*) \;\leq\; \frac{1}{2r}, \quad \text{and} \tag{26b}$$

$$\textbf{(A3)} \quad \frac{\sigma_m}{\sqrt{n}} \mathcal{F}(t) \;\leq \sqrt{\kappa} t \quad \text{for the same constant } \kappa \text{ as in (A1).} \tag{26c}$$

$$\textbf{(A4)} \quad r \leq \min\Big\{\frac{m}{2}, \frac{n}{4}, \kappa\frac{\sqrt{n}}{\sigma_m}\Big\}. \tag{26d}$$

Remarks: The first part of condition (A1) is to prevent the ratio s_r/s_1 from going to zero as the pair (m, n) increases, where the constant 1/2 is chosen for convenience. Such a lower bound is necessary for consistent estimation of the eigen-subspace corresponding to $\{s_1, \ldots, s_r\}$. The second part of condition (A1), involving the constant κ, provides a lower bound on the signal-to-noise ratio s_r/σ_m. Condition (A2) is required to prevent degeneracy among the vectors $z_j^* = \Phi f_j^*$ obtained by mapping the unknown eigenfunctions to the observation space $\mathbb{R}^m$. (In the ideal setting, we would have $C_m(f^*) = 0$, but our analysis shows that the upper bound in (A2) is sufficient.) Condition (A3) is required so that the critical tolerance $\epsilon_{m,n}$ specified below is well-defined; as will be clarified, it is always satisfied for the time-sampling model, and holds for the basis truncation model whenever $n \geq m$. Condition (A4) is easily satisfied, since the RHS of (26d) goes to ∞ while we usually take r to be fixed. Our results, however, hold if r grows slowly with m and n subject to (26d).

Theorem 1. *Under conditions (A1)—(A3) for a sufficiently small constant κ, let $\epsilon_{m,n}$ be the smallest positive number satisfying the inequality*

$$\frac{\sigma_m}{\sqrt{n}} r^{3/2} \mathcal{F}(\epsilon) \leq \kappa\, \epsilon^2. \tag{27}$$

Then there are universal positive constants (c_0, c_1, c_2) such that

$$\mathbb{P}\big[\, \mathrm{d}_{HS}^2(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \leq c_0\, \epsilon_{m,n}^2 \big] \geq 1 - \varphi(n, \epsilon_{m,n}), \tag{28}$$

where $\varphi(n, \epsilon_{m,n}) := c_1\big\{r^2 \exp\big(-c_2 r^{-3} \frac{n}{\sigma_m^2}(\epsilon_{m,n} \wedge \epsilon_{m,n}^2)\big) + r\exp(-\frac{n}{64})\big\}$.

We note that Theorem 1 is a general result, applying to an arbitrary bounded linear operator Φ. However, we can obtain a number of concrete results by making specific choices of this sampling operator, as we explore in the following sections.

4.1.1 Consequences for time-sampling

Let us begin with the time-sampling model (8), in which we observe the sampled functions

$$y_i = \begin{bmatrix} x_i(t_1) & x_i(t_2) & \ldots & x_i(t_m) \end{bmatrix}^T + \sigma_0 w_i, \quad \text{for } i = 1, 2, \ldots, m.$$

As noted earlier, this set-up can be modeled in our general setting (10) with $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$ and $\sigma_m = \sigma_0/\sqrt{m}$.

In this case, by the reproducing property of the RKHS, the matrix $K = \Phi\Phi^*$ has entries of the form $K_{ij} = \langle \phi_i, \phi_j \rangle_{\mathcal{H}} = \frac{\mathbb{K}(t_i,t_j)}{m}$. Letting $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \ldots \geq \widehat{\mu}_m > 0$ denote its ordered eigenvalues, we say that the kernel matrix K has polynomial-decay with parameter $\alpha > 1/2$ if there is a constant c such that $\widehat{\mu}_j \leq c\, j^{-2\alpha}$ for all $j = 1, 2, \ldots, m$. Since the kernel matrix K represents a discretized approximation of the kernel integral operator defined by $\mathbb{K}$, this type of polynomial decay is to be expected whenever the kernel operator has polynomial-α decaying eigenvalues. For example, the usual spline kernels that define Sobolev spaces have this type of polynomial decay [14]. In Appendix A, we verify this property explicitly for the kernel $\mathbb{K}(s,t) = \min\{s,t\}$ that defines the Sobolev class with smoothness $\alpha = 1$.

For any such kernel, we have the following consequence of Theorem 1:

Corollary 1 (Achievable rates for time-sampling)**.** *Consider the case of a time-sampling operator Φ. In addition to conditions (A1) and (A2), suppose that the kernel matrix K has polynomial-decay with parameter $\alpha > 1/2$. Then we have*

$$\mathbb{P}\Big[\mathrm{d}^2_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \leq c_0 \min\big\{\big(\frac{\kappa_{r,\rho}\,\sigma_0^2}{mn}\big)^{\frac{2\alpha}{2\alpha+1}},\; r^3\,\frac{\sigma_0^2}{n}\big\}\Big] \geq 1 - \varphi(n,m), \tag{29}$$

where $\kappa_{r,\rho} := r^{3+\frac{1}{2\alpha}}\,\rho^{\frac{1}{\alpha}}$, and $\varphi(n,m) := c_1\big\{\exp\big(-c_2\big\{\big(r^{-2}\rho^2 mn\big)^{\frac{1}{2\alpha+1}} \wedge m\big\}\big) + \exp(-n/64)\big\}$.

Remarks: (a) Disregarding constant pre-factors not depending on the pair (m,n), Corollary 1 guarantees that solving the program (19) returns a subspace estimate $\widehat{\mathfrak{Z}}$ such that

$$\mathrm{d}^2_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \precsim \min\big\{(mn)^{-\frac{2\alpha}{2\alpha+1}},\, n^{-1}\big\} \qquad \text{with high probability as } (m,n) \text{ increase.}$$

Depending on the scaling of the number of time samples m relative to the number of functional samples n, either term in this upper bound can be the smallest (and hence active) one. For instance, it can be verified that whenever $m \geq n^{\frac{1}{2\alpha}}$, then the first term is smallest, so that we achieve the rate $\mathrm{d}^2_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \precsim (mn)^{-\frac{2\alpha}{2\alpha+1}}$. The appearance of the term $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ is quite natural, as it corresponds to the minimax rate of a non-parameteric regression problem with smoothness α, based on m samples each of variance n^{-1}. Later, in Section 4.3, we provide results guaranteeing that this scaling is minimax optimal under reasonable conditions on the choice of sample points (in particular, see Theorem 3(a)).

(b) To be clear, although the bound (29) allows for the possibility that the error is of order *lower than* n^{-1}, we note that the probability with which the guarantee holds includes a term of the order $\exp(-n/64)$. Consequently, in terms of expected error, we cannot guarantee a rate faster than n^{-1}.

Proof. We need to bound the critical value $\epsilon_{m,n}$ defined in the theorem statement (27). Define the function $\mathcal{G}^2(t) := \sum_{j=1}^m \min\{\widehat{\mu}_j, t^2\}$, and note that $\mathcal{F}(t) = \sqrt{r}\rho\,\mathcal{G}(\frac{t}{\sqrt{r}\rho})$ by construction. Under the assumption of polynomial-α eigendecay, we have

$$\mathcal{G}^2(t) \leq \int_0^\infty \min\{cx^{-2\alpha}, t^2\}\, dx,$$

and some algebra then shows that $\mathcal{G}(t) \precsim t^{1-1/(2\alpha)}$. Disregarding constant factors, an upper bound on the critical $\epsilon_{m,n}$ can be obtained by solving the equation

$$\epsilon^2 = \frac{\sigma_m}{\sqrt{n}} r^{3/2} \sqrt{r}\rho \big(\frac{\epsilon}{\sqrt{r}\rho}\big)^{1-1/(2\alpha)}.$$

Doing so yields the upper bound $\epsilon^2 \precsim \big[\frac{\sigma_m^2}{n} r^3 (\sqrt{r}\rho)^{\frac{1}{\alpha}}\big]^{\frac{2\alpha}{2\alpha+1}}$. Otherwise, we also have the trivial upper bound $\mathcal{F}(t) \leq \sqrt{m}\, t$, which yields the alternative upper bound $\varepsilon_{m,n} \precsim \big(\frac{m\,\sigma_m^2}{n} r^3\big)^{1/2}$. Recalling that $\sigma_m = \sigma_0/\sqrt{m}$ and combining the pieces yields the claim. Notice that this last (trivial) bound on $\mathcal{F}(t)$ implies that condition (A3) is always satisfied for the time-sampling model. □

4.1.2 Consequences for basis truncation

We now turn to some consequences for the basis truncation model (9).

Corollary 2 (Achievable rates for basis truncation)**.** *Consider a basis truncation operator Φ in a Hilbert space with polynomial-α decay. Under conditions (A1), (A2) and $m \leq n$, we have*

$$\mathbb{P}\big[\, \mathrm{d}_{HS}^2(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \leq c_0 \,\big(\frac{\kappa_{r,\rho}\,\sigma_0^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}}\big] \geq 1 - \varphi(n,m), \tag{30}$$

where $\kappa_{r,\rho} := r^{3+\frac{1}{2\alpha}}\, \rho^{\frac{1}{\alpha}}$, and $\varphi(n,m) := c_1\big\{\exp\big(-c_2\big(r^{-2}\rho^2 n\big)^{\frac{1}{2\alpha+1}}\big) + \exp(-n/64)\big\}$.

Proof. We note that as long as $m \leq n$, condition (A3) is satisfied, since $\frac{\sigma_m}{\sqrt{n}}\mathcal{F}(t) \leq \sigma_0\sqrt{\frac{m}{n}} t \leq \sigma_0 t$. The rest of the proof follows that of Corollary 1, noting that in the last step we have $\sigma_m = \sigma_0$ for the basis truncation model. □

4.2 Function-based estimation rates (for $\widehat{\mathfrak{F}}$)

As mentioned earlier, given the consistency of $\widehat{\mathfrak{Z}}$, the consistency of $\widehat{\mathfrak{F}}$ is closely related to approximation properties of the semi-norm $\|\cdot\|_\Phi$ induced by Φ, and in particular how closely it approximates the L^2-norm. These approximation-theoretic properties are captured in part by the nullspace width $N_m(\Phi)$ and defect $D_m(\Phi)$ defined earlier in equations (13) and (15) respectively. In addition to these previously defined quantities, we require bounds on the following global quantity

$$R_m(\epsilon;\nu) := \sup\big\{\|f\|_{L^2}^2 \;\mid\; \|f\|_{\mathcal{H}}^2 \leq \nu^2,\ \|f\|_\Phi^2 \leq \epsilon^2\big\}. \tag{31}$$

A general upper bound on this quantity is of the form

$$R_m(\epsilon;\nu) \leq c_1\epsilon^2 + \nu^2 S_m(\Phi). \tag{32}$$

In fact, it is not hard to show that such a bound exists with $c_1 = 2$ and $S_m(\Phi) = 2(D_m(\Phi) + N_m(\Phi))$ using the decomposition $\mathcal{H} = \mathrm{Ra}(\Phi^*) \oplus \mathrm{Ker}(\Phi)$. However, this bound is not sharp. Instead, one can show that in most cases of interest, the term $S_m(\Phi)$ is of the order of $N_m(\Phi)$.

There are a variety of conditions that ensure that $S_m(\Phi)$ has this scaling; we refer the reader to the paper [2] for a general approach. Here we provide a simple sufficient condition, namely:

$$\textbf{(B1)} \qquad \Theta \preceq c_0 K^2 \tag{33}$$

for a positive constant c_0.

Lemma 5. *Under (B1), the bound (32) holds with $c_1 = 2c_0$ and $S_m(\Phi) = 2N_m(\Phi)$.*

See Appendix B.4 for the proof of this claim. In the sequel, we show that the first-order Sobolev RKHS satisfies the condition **(B1)**.

Theorem 2. *Suppose that condition (A1) holds, and the approximation-theoretic quantities satisfy the bounds $D_m(\Phi) \leq \frac{1}{4r\rho^2} \leq 1$ and $N_m(\Phi) \leq 1$. Then there is a constant $\kappa'_{r,\rho}$ such that*

$$\mathrm{d}_{HS}^2(\widehat{\mathfrak{F}}, \mathfrak{F}^*) \leq \kappa'_{r,\rho}\{\epsilon_{m,n}^2 + S_m(\Phi) + [D_m(\Phi)]^2\} \tag{34}$$

with the same probability as in Theorem 1.

As with Theorem 1, this is a generally applicable result, stated in abstract form. By specializing it to different sampling models, we can obtain concrete rates, as illustrated in the following sections.

4.2.1 Consequences for time-sampling

We begin by returning to the case of the time sampling model (8), where $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$. In this case, condition (B1) needs to be verified by some calculations. For instance, as shown in Appendix A, in the case of the Sobolev kernel with smoothness $\alpha = 1$ (namely, $\mathbb{K}(s,t) = \min\{s,t\}$), we are guaranteed that (B1) holds with $c_0 = 1$, whenever the samples $\{t_j\}$ are chosen uniformly over $[0,1]$; hence, by Lemma 5, $S_m(\Phi) = 2N_m(\Phi)$. Moreover, in the case of uniform sampling, we expect that the nullspace width $N_m(\Phi)$ is upper bounded by μ_{m+1}, so will be proportional to $m^{-2\alpha}$ in the case of a kernel operator with polynomial-α decay. This is verified in [2] (up to a logarithmic factor) for the case of the first-order Sobolev kernel. In Appendix A, we also show that, for this kernel, $[D_m(\Phi)]^2$ is of the order $m^{-2\alpha}$, that is, of the same order as $N_m(\Phi)$.

Corollary 3. *Consider the basis truncation model* (9) *with uniformly spaced samples, and assume condition (B1) holds and that $N_m(\Phi) + [D_m(\Phi)]^2 \precsim m^{-2\alpha}$. Then the M-estimator returns a subspace estimate $\widehat{\mathfrak{F}}$ such that*

$$\mathrm{d}_{HS}^2(\widehat{\mathfrak{F}}, \mathfrak{F}^*) \leq \kappa'_{r,\rho}\Big\{\min\{\big(\frac{\sigma_0^2}{nm}\big)^{\frac{2\alpha}{2\alpha+1}}, \frac{\sigma_0^2}{n}\} + \frac{1}{m^{2\alpha}}\Big\} \tag{35}$$

with the same probability as in Corollary 1.

In this case, there is an interesting trade-off between the bias or approximation error terms which is of order $m^{-2\alpha}$ and the estimation error. An interesting transition occurs at the point when $m \succsim n^{\frac{1}{2\alpha}}$, at which:

- the bias term $m^{-2\alpha}$ becomes of the order n^{-1}, so that it is no longer dominant, and
- for the two terms in the estimation error, we have the ordering

$$(mn)^{-\frac{2\alpha}{2\alpha+1}} \leq \big(n^{1+\frac{1}{2\alpha}}\big)^{-\frac{2\alpha}{2\alpha+1}} = n^{-1}.$$

Consequently, we conclude that the scaling $m = n^{\frac{1}{2\alpha}}$ is the minimal number of samples such that we achieve an overall bound of the order n^{-1} in the time-sampling model. In Section 4.3, we will see that these rates are minimax-optimal.

4.2.2 Consequences for basis truncation

For the basis truncation operator Φ, we have $\Theta = K^2 = \mathrm{diag}(\mu_1^2, \dots, \mu_m^2)$ so that condition (B1) is satisfied trivially with $c_0 = 1$. Moreover, Lemma 2 implies $D_m(\Phi) = 0$. In addition, a function $f = \sum_{j=1}^{\infty} \sqrt{\mu_j} a_j \psi_j$ satisfies $\Phi f = 0$ if and only if $a_1 = a_2 = \dots = a_m = 0$, so that

$$N_m(\Phi) = \sup\bigl\{\|f\|_{L^2}^2 \mid \|f\|_{\mathcal{H}} \le 1,\ \Phi f = 0\bigr\} \;=\; \mu_{m+1}.$$

Consequently, we obtain the following corollary of Theorem 2:

Corollary 4. *Consider the basis truncation model* (9) *with a kernel operator that has polynomial-α decaying eigenvalues. Then the M-estimator returns a function subspace estimate $\widehat{\mathfrak{F}}$ such that*

$$\mathrm{d}_{HS}^2(\widehat{\mathfrak{F}}, \mathfrak{F}^*) \le \kappa'_{r,\rho}\bigl\{\bigl(\frac{\sigma_0^2}{n}\bigr)^{\frac{2\alpha}{2\alpha+1}} + \frac{1}{m^{2\alpha}}\bigr\} \tag{36}$$

with the same probability as in Corollary 2.

By comparison to Corollary 3, we see that the trade-offs between (m, n) are very different for basis truncation. In particular, there is *no interaction* between the number of functional samples m and the number of statistical samples n. Increasing m only reduces the approximation error, whereas increasing n only reduces the estimation error. Moreover, in contrast to the time sampling model of Corollary 3, it is impossible to achieve the fast rate n^{-1}, regardless of how we choose the pair (m, n). In Section 4.3, we will also see that the rates given in Corollary 4 are minimax optimal.

4.3 Lower bounds

We now turn to lower bounds on the minimax risk, demonstrating the sharpness of our achievable results in terms of their scaling with (m, n). In order to do so, it suffices to consider the simple model with a single functional component $f^* \in \mathbb{B}_{\mathcal{H}}(1)$, so that we observe $y_i = \beta_{i1}\,\Phi_m(f^*) + \sigma_m w_i$ for $i = 1, 2, \dots, n$, where $\beta_{i1} \sim N(0, 1)$ are i.i.d. standard normal variates. The minimax risk over the unit ball of the function space $\mathcal{H}$ in the Φ-norm is given by

$$\mathcal{M}_{m,n}^{\mathcal{H}}(\|\cdot\|_{\Phi}) := \inf_{\widetilde{f}} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}(1)} \mathbb{E}\|\widetilde{f} - f^*\|_{\Phi}^2, \tag{37}$$

where the function f^* ranges over the unit ball $\mathbb{B}_{\mathcal{H}}(1) = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \le 1\}$ of some Hilbert space, and $\widetilde{f}$ ranges over measurable functions of the data matrix $(y_1, y_2, \dots, y_n) \in \mathbb{R}^{m \times n}$.

Theorem 3 (Lower bounds for $\|\widetilde{f} - f^*\|_{\Phi}$)**.** *Suppose that the kernel matrix K has eigenvalues with polynomial-α decay and (A1) holds.*

(a) For the time-sampling model, we have

$$\mathcal{M}_{m,n}^{\mathcal{H}}(\,\|\cdot\|_{\Phi}) \ge C \min\bigl\{\,\bigl(\frac{\sigma_0^2}{mn}\bigr)^{\frac{2\alpha}{2\alpha+1}},\ \frac{\sigma_0^2}{n}\bigr\}. \tag{38}$$

(b) For the frequency-truncation model, with $m \ge (c_0 n)^{\frac{1}{2\alpha+1}}$, we have:

$$\mathcal{M}_{m,n}^{\mathcal{H}}(\|\cdot\|_{\Phi}) \ge C\,\bigl(\frac{\sigma_0^2}{n}\bigr)^{\frac{2\alpha}{2\alpha+1}}. \tag{39}$$

Note that part (a) of Theorem 3 shows that the rates obtained in Corollary 3 for the case of time-sampling are minimax optimal. Similarly, comparing part (b) of the theorem to Corollary 4, we conclude that the rates obtained for frequency truncation model are minimax optimal for $n \in [m, c_1 m^{2\alpha+1}]$. As will become clear momentarily (as a consequence of our next theorem), the case $n > c_1 m^{2\alpha+1}$ is not of practical interest.

We now turn to lower bounds on the minimax risk in the $\|\cdot\|_{L^2}$ norm—namely

$$\mathcal{M}^{\mathcal{H}}_{m,n}(\delta^2;\ \|\cdot\|_{L^2}) := \inf_{\widetilde{f}} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}(1)} \mathbb{E}\|\widetilde{f} - f^*\|^2_{L^2}. \tag{40}$$

Obtaining lower bounds on this minimax risk requires another approximation property of the norm $\|\cdot\|_\Phi$ relative to $\|\cdot\|_{L^2}$. Recall the the matrix $\Psi \in \mathbb{R}^{m\times m}$ with entries $\Psi_{ij} := \langle \psi_i, \psi_j \rangle_\Phi$. Since the eigenfunctions are orthogonal in L^2, the deviation of Ψ from the identity measures how well the inner product defined by Φ approximates the L^2-inner product over the first m eigenfunctions of the kernel operator. For proving lower bounds, we require an upper bound of the form

$$\textbf{(B2)} \qquad \lambda_{\max}(\Psi) \leq c_1,$$

for some universal constant $c_1 > 0$. As the proof will clarify, this upper bound is necessary in order that the Kullback-Leibler divergence—-which controls the relative discriminability between different models—can be upper bounded in terms of the L^2-norm.

Theorem 4 (Lower bounds for $\|\widetilde{f} - f^*\|^2_{L^2}$)**.** *Suppose that condition (B2) holds, and the operator associated with kernel function $\mathbb{K}$ of the reproducing kernel Hilbert space $\mathcal{H}$ has eigenvalues with polynomial-α-decay.*

(a) For the time-sampling model, the minimax risk is lower bounded as

$$\mathcal{M}^{\mathcal{H}}_{m,n}(\|\cdot\|_{L^2}) \geq\ C\left\{ \min\{(\frac{\sigma_0^2}{mn})^{\frac{2\alpha}{2\alpha+1}},\ \frac{\sigma_0^2}{n}\} + (\frac{1}{m})^{2\alpha}\right\}. \tag{41}$$

(b) For the frequency-truncation model, the minimax error is lower bounded as

$$\mathcal{M}^{\mathcal{H}}_{m,n}(\|\cdot\|_{L^2}) \geq C\left\{ (\frac{\sigma_0^2}{n})^{\frac{2\alpha}{2\alpha+1}} + (\frac{1}{m})^{2\alpha}\right\}. \tag{42}$$

Verifying condition (B2) requires, in general, some calculations in the case of time-sampling model. It is verified for uniform time-sampling for the first-order Sobolev RKHS in Appendix A. For the frequency-truncation model, condition (B2) always holds trivially since $\Psi = I_m$. By this theorem, the L^2 convergence rates of Corollary 3 and 4 are minimax optimal. Also note that due to the presence of the approximation term $m^{-2\alpha}$ in (42), the Φ-norm term $n^{\frac{2\alpha}{2\alpha+1}}$ is only dominant when $m \geq c_2 n^{\frac{1}{2\alpha+1}}$ implying that this is the interesting regime for Theorem 3(b).

5 Proof of subspace-based rates

We now turn to the proofs of the results involving the error $\mathrm{d}_{HS}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*)$ between the estimated $\widehat{\mathfrak{Z}}$ and true subspace $\mathfrak{Z}^*$. We begin by proving Theorem 1, and then turn to its corollaries.

5.1 Preliminaries

We begin with some preliminaries before proceeding to the heart of the proof. Let us first introduce some convenient notation. Consider the $n \times m$ matrices

$$Y := \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T, \quad \text{and} \quad W := \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}^T,$$

corresponding to the observation matrix Y and noise matrix W respectively. In addition, we define the matrices $B := (\beta_{ij}) \in \mathbb{R}^{n \times r}$ and $S := \operatorname{diag}(s_1, \ldots, s_r) \in \mathbb{R}^{r \times r}$. Recalling that $Z^* := (z_1^*, \ldots, z_r^*) \in \mathbb{R}^{m \times r}$, the observation model (10) can be written in the matrix form $Y = B(Z^* S)^T + \sigma_m W$. Moreover, let us define the matrices $\overline{B} := \frac{B^T B}{n} \in \mathbb{R}^{r \times r}$ and $\overline{W} := \frac{W^T B}{n} \in \mathbb{R}^{m \times r}$. Using this notation, some algebra shows that the associated sample covariance $\widehat{\Sigma}_n := \frac{1}{n} Y^T Y$ can be written in the form

$$\widehat{\Sigma}_n = \underbrace{Z^* S \overline{B} S (Z^*)^T}_{\Gamma} + \Delta_1 + \Delta_2, \tag{43}$$

where $\Delta_1 := \sigma_m \big[\overline{W} S (Z^*)^T + Z^* S \overline{W}^T\big]$ and $\Delta_2 := \sigma_m^2 \frac{W^T W}{n}$.

Lemma 3, proved in Appendix B.3 shows the existence of a matrix $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$ such that $\operatorname{Ra}(\widetilde{Z}^*) = \operatorname{Ra}(Z^*)$. As discussed earlier, due to the nature of the Steifel manifold, there are many versions of this matrix $\widetilde{Z}^*$, and also of any optimal solution matrix $\widehat{Z}$, obtained via right multiplication with an orthogonal matrix. For the subsequent arguments, we need to work with a particular version of $\widetilde{Z}^*$ (and $\widehat{Z}$) that we describe here.

Now let us fix some convenient versions of $\widetilde{Z}^*$ and $\widehat{Z}$. As a consequence of CS decomposition, as long as $r \leq m/2$, there exist orthogonal matrices $U, V \in \mathbb{R}^{r \times r}$ and an orthogonal matrix $Q \in \mathbb{R}^{m \times m}$, such that

$$Q^T \widetilde{Z}^* U = \begin{pmatrix} I_r \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad Q^T \widehat{Z} V = \begin{pmatrix} \widehat{C} \\ \widehat{S} \\ 0 \end{pmatrix}, \tag{44}$$

where $\widehat{C} = \operatorname{diag}(\widehat{c}_1, \cdots, \widehat{c}_r)$ and $\widehat{S} = \operatorname{diag}(\widehat{s}_1, \cdots, \widehat{s}_r)$ such that $1 \geq \widehat{s}_1 \geq \cdots \geq \widehat{s}_r \geq 0$ and $\widehat{C}^2 + \widehat{S}^2 = I_r$. (See Bhatia [5], Theorem VII.1.8, for details on this decomposition.) In the analysis to follow, we work with $\widetilde{Z}^* U$ and $\widehat{Z} V$ instead of $\widetilde{Z}^*$ and $\widehat{Z}$. To avoid extra notation, from now on, we will use $\widetilde{Z}^*$ and $\widehat{Z}$ for these new versions, which we refer to as *properly aligned*. With this choice, we may assume $U = V = I_r$ in the CS decomposition (44).

The following lemma isolates some useful properties of properly aligned subspaces:

Lemma 6. *Let $\widetilde{Z}^*$ and $\widehat{Z}$ be properly aligned, and define the matrices*

$$\widehat{P} := P_{\widehat{Z}} - P_{\widetilde{Z}^*} \;=\; \widehat{Z}\widehat{Z}^T - \widetilde{Z}^* (\widetilde{Z}^*)^T, \quad \text{and} \quad \widehat{E} := \widehat{Z} - \widetilde{Z}^*. \tag{45}$$

In terms of the CS decomposition (44), we have

$$|||\widehat{E}|||_{HS} \leq |||\widehat{P}|||_{HS}, \tag{46a}$$

$$(\widetilde{Z}^*)^T (P_{\widetilde{Z}^*} - P_{\widehat{Z}}) \widetilde{Z}^* = \widehat{S}^2, \quad \text{and} \tag{46b}$$

$$d_{HS}^2(\widehat{Z}, \widetilde{Z}^*) = |||P_{\widetilde{Z}^*} - P_{\widehat{Z}}|||_{HS}^2 = 2|||\widehat{S}^2|||_{HS}^2 + 2|||\widehat{C}\widehat{S}|||_{HS}^2 = 2\sum_k \widehat{s}_k^2(\widehat{s}_k^2 + \widehat{c}_k^2) = 2\operatorname{tr}(\widehat{S}^2). \tag{46c}$$

Proof. From the CS decomposition (44), we have $\widetilde{Z}^*(\widetilde{Z}^*)^T - \widehat{Z}(\widehat{Z})^T = Q\begin{pmatrix} \widehat{S}^2 & -\widehat{C}\widehat{S} & 0 \\ -\widehat{S}\widehat{C} & -\widehat{S}^2 & 0 \\ 0 & 0 & 0 \end{pmatrix}Q^T$, from which relations (46b) and (46c) follow. From the decomposition (44) and the proper alignment condition $U = V = I_r$, we have

$$\begin{aligned} |||\widehat{E}|||_{HS}^2 = |||Q^T(\widehat{Z} - \widetilde{Z}^*)|||_{HS}^2 &= |||I_r - \widehat{C}|||_{HS}^2 + |||\widehat{S}|||_{HS}^2 \\ &= 2\sum_{i=1}^r (1 - \widehat{c}_i) \leq 2\sum_{i=1}^r (1 - \widehat{c}_i^2) = 2\sum_{i=1}^r \widehat{s}_i^2 = |||\widehat{P}|||_{HS}^2 \end{aligned} \qquad (47)$$

where we have used the relations $\widehat{C}^2 + \widehat{S}^2 = I_r$, $\widehat{c}_i \in [0, 1]$, and $2\,\mathrm{tr}(\widehat{S}^2) = |||P_{\widetilde{Z}^*} - P_{\widehat{Z}}|||_{HS}^2$. □

5.2 Proof of Theorem 1

Using the notation introduced in Lemma 6, our goal is to bound the Hilbert-Schmidt norm $|||\widehat{P}|||_{HS}$. Without loss of generality we will assume $s_1 = 1$ throughout. Recalling the definition (43) of the random matrix Δ, the following inequality plays a central role in the proof:

Lemma 7. *Under condition (A1) and* $s_1 = 1$, *we have*

$$|||\widehat{P}|||_{HS}^2 \leq 128 \,\langle\!\langle \widehat{P},\ \Delta_1 + \Delta_2 \rangle\!\rangle \qquad (48)$$

with probability at least $1 - \exp(-n/32)$.

Proof. We use the shorthand notation $\Delta = \Delta_1 + \Delta_2$ for the proof. Since $\widetilde{Z}^*$ is feasible and $\widehat{Z}$ is optimal for the program (19), we have the basic inequality $\langle\!\langle \widehat{\Sigma}_n,\ P_{\widetilde{Z}^*} \rangle\!\rangle \leq \langle\!\langle \widehat{\Sigma}_n,\ P_{\widehat{Z}} \rangle\!\rangle$. Using the decomposition $\widehat{\Sigma} = \Gamma + \Delta$ and rearranging yields the inequality

$$\langle\!\langle \Gamma,\ P_{\widetilde{Z}^*} - P_{\widehat{Z}} \rangle\!\rangle \leq \langle\!\langle \Delta,\ P_{\widehat{Z}} - P_{\widetilde{Z}^*} \rangle\!\rangle. \qquad (49)$$

From the definition (43) of Γ and $Z^* = \widetilde{Z}^* R$, the left-hand side of the inequality (49) can be lower bounded as

$$\begin{aligned} \langle\!\langle \Gamma,\ P_{\widetilde{Z}^*} - P_{\widehat{Z}} \rangle\!\rangle &= \langle\!\langle \overline{B},\ SR^T(\widetilde{Z}^*)^T(P_{\widetilde{Z}^*} - P_{\widehat{Z}})\widetilde{Z}^* RS \rangle\!\rangle \\ &= \mathrm{tr}\,\overline{B}SR^T\widehat{S}^2 RS \\ &\geq \lambda_{\min}(\overline{B})\lambda_{\min}(S^2)\lambda_{\min}(R^T R)\,\mathrm{tr}(\widehat{S}^2) \end{aligned}$$

where we have used (89) and (90) of Appendix I, several times. We note that $\lambda_{\min}(S^2) = s_r^2 \geq \frac{1}{2}$ and $\lambda_{\min}(R^T R) \geq \frac{1}{2}$ provided $rC_m(f^*) \geq \frac{1}{2}$; see equation (69). To bound the minimum eigenvalue of $\overline{B}$, let $\gamma_{\min}(B)$ denote the minimum singular value of the $n \times r$ Gaussian matrix B. The following concentration inequality is well-known (cf. [11, 20]):

$$\mathbb{P}\big[\gamma_{\min}(B) \leq \sqrt{n} - \sqrt{r} - t\big] \leq \exp(-t^2/2), \quad \text{for all } t > 0.$$

Since $\lambda_{\min}(\overline{B}) = \gamma_{\min}^2(B/\sqrt{n})$, we have that $\lambda_{\min}(\overline{B}) \geq (1 - \sqrt{r/n} - t)^2$ with probability at least $1 - \exp(-nt^2/2)$. Assuming $r/n \leq \frac{1}{4}$ and setting $t = \frac{1}{4}$, we get $\lambda_{\min}(\overline{B}) \geq \frac{1}{16}$ with probability at least $1 - \exp(-n/32)$. Putting the pieces together yields the claim. □

The inequality (48) reduces the problem of bounding $\|\widehat{P}\|_{HS}^2$ to the sub-problem of studying the random variable $\langle\!\langle \widehat{P},\ \Delta_1 + \Delta_2 \rangle\!\rangle$. Based on Lemma 7, our next step is to establish an inequality (holding with high probability) of the form

$$\langle\!\langle \widehat{P},\ \Delta_1 + \Delta_2 \rangle\!\rangle \leq c_1 \{ \frac{\sigma_m}{\sqrt{n}} r^{3/2} \mathcal{F}(\|\widehat{E}\|_{HS}) + \kappa \|\widehat{E}\|_{HS}^2 + \epsilon_{m,n}^2 \}, \tag{50}$$

where c_1 is some universal constant, κ is the constant in condition (A1), and $\epsilon_{m,n}$ is the critical radius from Theorem 1. Doing so is a non-trivial task: both matrices $\widehat{P}$ and Δ are random and depend on one another, since the subspace $\widehat{Z}$ was obtained by optimizing a random function depending on Δ. Consequently, our proof of the bound (50) involves deriving a uniform law of large numbers for a certain matrix class.

Suppose that the bound (50) holds, and that the subspaces $\widetilde{Z}^*$ and $\widehat{Z}$ are properly aligned. Lemma 6 implies that $\|\widehat{E}\|_{HS} \leq \|\widehat{P}\|_{HS}$, and since $\mathcal{F}$ is a non-decreasing function, the inequality (50) combined with Lemma 7 implies that

$$\big(1 - 128\kappa c_1\big) \|\widehat{P}\|_{HS}^2 \leq c_1 \Big\{ \frac{\sigma_m}{\sqrt{n}} r^{3/2} \mathcal{F}(\|\widehat{P}\|_{HS}) + \epsilon_{m,n}^2 \Big\},$$

from which the claim follows as long as κ is suitably small (for instance, $\kappa \leq \frac{c_1}{256}$ suffices). Accordingly, in order to complete the proof of Theorem 1, it remains to prove the bound (50), and the remainder of our work is devoted to this goal. Given the linearity of trace, we can bound the terms $\langle\!\langle \widehat{P},\ \Delta_1 \rangle\!\rangle$ and $\langle\!\langle \widehat{P},\ \Delta_2 \rangle\!\rangle$ separately.

5.2.1 Bounding $\langle\!\langle \widehat{P},\ \Delta_1 \rangle\!\rangle$

Let $\{\overline{z}_j\}$, $\{\widetilde{z}_j^*\}$ and $\{\widehat{e}_j\}$ and $\{\overline{w}_j\}$ denote the columns of $\widehat{Z}$, $\widetilde{Z}^*$, $\widehat{E}$ and $\overline{W}$, respectively, where we recall the definitions of these quantities from equation (43) and Lemma 6. Note that $\overline{w}_j = n^{-1} \sum_{i=1}^n w_i \beta_{ij}$. In Appendix C.1, we show that

$$\langle\!\langle \widehat{P},\ \Delta_1 \rangle\!\rangle \leq \sqrt{6}\, \sigma r^{3/2} \max_{j,k} |\langle \overline{w}_k, \widehat{e}_j \rangle| + \sqrt{\frac{3}{2}}\, \sigma r\, \|\widehat{E}\|_{HS}^2 \max_{j,k} |\langle \overline{w}_j, \widetilde{z}_k^* \rangle|. \tag{51}$$

Consequently, we need to obtain bounds on quantites of the form $|\langle \overline{w}_j, v \rangle|$, where the vector v is either fixed (e.g., $v = \widetilde{z}_j^*$) or random (e.g., $v = \widehat{e}_j$). The following lemmas provide us with the requisite bounds:

Lemma 8. *We have*

$$\max_{j,k} \sigma r^{3/2}\ |\langle \overline{w}_k, \widehat{e}_j \rangle|\ \leq\ C \Big\{ \frac{\sigma}{\sqrt{n}} r^{3/2} \mathcal{F}(\|\widehat{E}\|_{HS}) + \kappa \|\widehat{E}\|_{HS}^2 + \kappa \epsilon_{m,n}^2 \Big\}$$

with probability at least $1 - c_1 r \exp(-\kappa^2 r^{-3} n \frac{\epsilon_{m,n}^2}{2\sigma^2}) - r \exp(-n/64)$.

Lemma 9. *We have*

$$\mathbb{P}\big[\max_{j,k} \sigma r |\overline{w}_k^T \widetilde{z}_j^*| \leq \sqrt{6}\kappa \big] \geq 1 - r^2 \exp(-\kappa^2 r^{-2} n / 2\sigma^2).$$

See Appendix C.2 and C.3, respectively, for the proofs of these claims.

5.2.2 Bounding $\langle\!\langle \widehat{P},\ \Delta_2 \rangle\!\rangle$

Recalling the definition (43) of Δ_2 and using linearity of the trace, we obtain

$$\langle\!\langle \widehat{P},\ \Delta_2 \rangle\!\rangle = \frac{\sigma^2}{n} \sum_{j=1}^{r} \Big\{ (\overline{z}_j)^T W^T W \overline{z}_j - (\widetilde{z}_j^*)^T W^T W \widetilde{z}_j^* \Big\}.$$

Since $\widehat{e}_j = \overline{z}_j - \widetilde{z}_j^*$, we have

$$\begin{aligned} \langle\!\langle \widehat{P},\ \Delta_2 \rangle\!\rangle &= \sigma^2 \sum_{j=1}^{r} \Big\{ 2(\widetilde{z}_j^*)^T \Big(\frac{1}{n} W^T W - I_r \Big) \widehat{e}_j + \frac{1}{n} \|W \widehat{e}_j\|_2^2 + 2(\widetilde{z}_j^*)^T \widehat{e}_j \Big\} \\ &\le \sigma^2 \sum_{j=1}^{r} \Big\{ \underbrace{2\, (\widetilde{z}_j^*)^T \Big(\frac{1}{n} W^T W - I_r \Big) \widehat{e}_j}_{T_1(\widehat{e}_j;\, \widetilde{z}_j^*)} + \underbrace{\frac{1}{n} \|W \widehat{e}_j\|_2^2}_{T_2(\widehat{e}_j)} \Big\}, \end{aligned} \tag{52}$$

where we have used the fact that $2\sum_j (\widetilde{z}_j^*)^T \widehat{e}_j = 2\sum_j [(\widetilde{z}_j^*)^T \overline{z}_j - 1] = 2\sum_j (\widehat{c}_j - 1) = -\|\!|\widehat{E}|\!\|_{HS}^2 \le 0$.

The following lemmas provide high probability bounds on the terms T_1 and T_2.

Lemma 10. *We have the upper bound*

$$\sigma^2 \sum_{j=1}^{r} T_1(\widehat{e}_j; \widetilde{z}_j^*) \le C \Big\{ \sigma_0 \frac{\sigma}{\sqrt{n}} r \mathcal{F}(\|\!|\widehat{E}|\!\|_{HS}) + \kappa \|\!|\widehat{E}|\!\|_{HS}^2 + \kappa \epsilon_{m,n}^2 \Big\}$$

with probability $1 - c_2 \exp(-\kappa^2 r^{-2} n \, \frac{\epsilon_{m,n} \wedge \epsilon_{m,n}^2}{16\sigma^2}) - r \exp(-n/64)$.

Lemma 11. *We have the upper bound* $\sigma^2 \sum_{j=1}^{r} T_2(\widehat{e}_j) \le C\kappa \big\{ \|\!|\widehat{E}|\!\|_{HS}^2 + \epsilon_{m,n}^2 \big\}$ *with probability at least* $1 - c_3 \exp(-\kappa^2 r^{-2} n \, \epsilon_{m,n}^2 / 2\sigma^2)$.

See Appendices C.4 and C.5, respectively, for the proofs of these claims.

6 Proof of functional rates

We now turn to the proof of Theorem 2, which provides upper bounds on the estimation error in the function domain. As in the proof of Theorem 1, let $\widehat{Z} = (\widehat{z}_1, \cdots, \widehat{z}_r) \in V_r(\mathbb{R}^m)$ and $\widetilde{Z}^* = (\widetilde{z}_1^*, \cdots, \widetilde{z}_r^*) \in V_r(\mathbb{R}^m)$ represent the subspaces $\widehat{\mathfrak{Z}}$ and $\mathfrak{Z}^*$ respectively, and assume that they are properly aligned (see Lemma 6). For $j = 1, \ldots, m$, define $\widehat{g}_j := \Phi^* K^{-1} \widehat{z}_j$ and $g_j^* := \Phi^* K^{-1} \widetilde{z}_j^*$. Let $\{\widehat{h}_j\}_{j=1}^r$ be *any* basis of $\widehat{\mathfrak{F}}$, orthonormal in L^2, and similarly, let $\{h_j^*\}_{j=1}^r$ be any orthonormal basis of $\mathfrak{F}^*$. Our goal is to bound the Hilbert-Schmidt norm $\|\!|P_{\widehat{\mathfrak{F}}} - P_{\mathfrak{F}^*}|\!\|_{\mathrm{HS}}^2$. In order to do so, we first observe that

$$\|\!|P_{\widehat{\mathfrak{F}}} - P_{\mathfrak{F}^*}|\!\|_{\mathrm{HS}}^2 \le 2 \sum_{j=1}^{r} \|\widehat{h}_j - h_j^*\|_{L^2}^2, \tag{53}$$

so that it suffices to upper bound $\sum_{j=1}^{r} \|\widehat{h}_j - h_j^*\|_{L^2}^2$. We relate this quantity to the functions $\widehat{g}_j$ and g_j^* via the elementary inequality

$$\|\widehat{h}_j - h_j^*\|_{L^2}^2 \le 4 \big\{ \|\widehat{g}_j - g_j^*\|_{L^2}^2 + \|\widehat{h}_j - \widehat{g}_j\|_{L^2}^2 + \|g_j^* - h_j^*\|_{L^2}^2 \big\}. \tag{54}$$

The remainder of our proof is focused on obtaining suitable upper bounds on each of these three terms.

We begin by bounding the first term $\|\widehat{g}_j - g_j^*\|_{L^2}^2$. Recall the definitions of $R_m(\epsilon;\nu)$ and $S_m(\Phi)$ and their relation via inequality (32). We exploit the inequality in the following way: suppose that we can show that

$$\sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_\Phi^2 \leq A^2, \quad \text{and} \quad \sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_{\mathcal{H}}^2 \leq B^2. \tag{55}$$

Let $S(A,B) = \{(a,b) \in \mathbb{R}^r \times \mathbb{R}^r \mid \sum_{j=1}^r a_j^2 \leq A^2,\ \sum_{j=1}^r b_j^2 \leq B^2\}$. We may then conclude that

$$\begin{aligned}\sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_{L^2}^2 &\leq \sup_{(a,b)\in S(A,B)} \sum_{j=1}^r R_m(a_j; b_j) \\ &\overset{(i)}{\leq} \sup_{(a,b)\in S(A,B)} \sum_{j=1}^r \left\{c_1 a_j^2 + b_j^2 S_m(\Phi)\right\} \\ &= c_1 A^2 + B^2 S_m(\Phi). \end{aligned} \tag{56}$$

where inequality (i) follows by repeated application of inequality (32).

It remains to establish upper bounds of the form (55). By definition, we have $\widehat{g}_j - g_j^* \in \mathrm{Ra}(\Phi^*)$ and $\Phi(\widehat{g}_j - g_j^*) = \widehat{z}_j - \widetilde{z}_j^*$. Recalling the norm $\|a\|_K^2 := a^T K^{-1} a$, we note that the matrices $\widehat{Z}$ and $\widetilde{Z}^*$ satisfy the trace smoothness condition $\sum_{j=1}^r \|\widehat{z}_j\|_K^2 = \langle\!\langle K^{-1},\ ZZ^T \rangle\!\rangle \leq 2r\rho^2$, and hence

$$\sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_{\mathcal{H}}^2 = \sum_{j=1}^r \|\widehat{z}_j - \widetilde{z}_j^*\|_K^2 \ \leq\ 2\sum_{j=1}^r (\|\widehat{z}_j\|_K^2 + \|\widetilde{z}_j^*\|_K^2) \ \leq\ \underbrace{8\,r\rho^2}_{B^2}$$

Furthermore, recalling that $\|f\|_\Phi = \|\Phi f\|_2$, we have

$$\sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_\Phi^2 = \sum_{j=1}^r \|\widehat{z}_j - \widetilde{z}_j^*\|_2^2 = |\!|\!|\widehat{Z} - \widetilde{Z}^*|\!|\!|_{HS}^2 \leq \underbrace{|\!|\!|P_{\widehat{Z}} - P_{\widetilde{Z}^*}|\!|\!|_{HS}^2}_{A^2}$$

Consequently, by the bound (56) with $A^2 = |\!|\!|P_{\widehat{Z}} - P_{\widetilde{Z}^*}|\!|\!|_{HS}^2$ and $B^2 = 8r\rho^2$, we conclude that

$$\sum_{j=1}^r \|\widehat{g}_j - g_j^*\|_{L^2}^2 \leq c_1 |\!|\!|P_{\widehat{\mathfrak{Z}}} - P_{\mathfrak{Z}^*}|\!|\!|_{HS}^2 + 8r\rho^2 S_m(\Phi) \tag{57}$$

We now need to bound the remaining two terms in the decomposition (54). In order to do so, we exploit the freedom in choosing the orthonormal families $\{\widehat{h}_j\}_{j=1}^r$ and $\{h_j^*\}_{j=1}^r$. By appropriate choices, we obtain the following results:

Lemma 12. *There exists an orthonormal basis $\{\widehat{h}_j\}_{j=1}^r$ of $\widehat{\mathfrak{F}}$ for which*

$$\sum_{j=1}^r \|\widehat{h}_j - \widehat{g}_j\|_{L^2}^2 = 2r^2\rho^4\, D_m^2(\Phi). \tag{58}$$

Lemma 13. *There exists an orthonormal basis $\{h_j^*\}_{j=1}^r$ of $\mathfrak{F}^*$ for which*

$$\sum_{j=1}^{r} \|h_j^* - g_j^*\|_{L^2}^2 \leq c_2\, r^2 C_m^2(f^*) + 6r\rho^2 S_m(\Phi). \tag{59}$$

As these proofs are more technical and lengthy, we defer them to Appendices D.1 and D.2 respectively.

Combining all of the pieces, we obtain the upper bound

$$|\!|\!| P_{\widehat{\mathfrak{F}}} - P_{\mathfrak{F}^*} |\!|\!|_{HS}^2 \leq c_3 \big\{ |\!|\!| P_{\widehat{\mathfrak{Z}}} - P_{\mathfrak{Z}^*} |\!|\!|_{HS}^2 + r^2\rho^4 D_m^2(\Phi) + r^2 C_m^2(f^*) + r\rho^2 S_m(\Phi) \big\}. \tag{60}$$

By using polarization identity and decomposition $\mathcal{H} = \mathrm{Ra}(\Phi^*) \oplus \mathrm{Ker}(\Phi)$, one can show that

$$C_m(f^*) \leq \kappa_\rho'' \big(D_m(\Phi) + N_m(\Phi) \big), \tag{61}$$

when $N_m(\Phi) \leq 1$. (See Appendix B.5 for more details.) Using this inequality and noting that $S_m(\Phi) \geq N_m(\Phi) \geq [N_m(\Phi)]^2$ when $N_m(\Phi) \leq 1$, the bound (60) can be simplified to the form given in Theorem 2.

7 Proof of minimax lower bounds

We now turn to the proofs of the minimax lower bounds stated in Theorems 3 and 4. We begin with some preliminary results that apply to both proofs.

7.1 Preliminary results

Our proofs proceed via a standard reduction from estimation to multi-way hypothesis testing (e.g., [35, 33]). In particular, let $\{f^1, \dots, f^M\}$ be an δ-packing set of $\mathbb{B}_{\mathcal{H}}(1)$ in a given norm $\|\cdot\|_\star$. (For our proofs, this norm will be either $\|\cdot\|_\Phi$ or $\|\cdot\|_{L^2}$.) Given such a packing set, it is known that the minimax error in the norm $\|\cdot\|_\star$ can be lower bounded, using Fano's inequality, by

$$\inf_{\widetilde{f}} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}(1)} \mathbb{P}_{f^*}\Big[\|\widetilde{f} - f^*\|_\star^2 \geq \frac{\delta^2}{4} \Big] \geq 1 - \frac{I(y; f) + \log 2}{\log M}. \tag{62}$$

where $y = (y_1, \dots, y_n) \in \mathbb{R}^{m \times n}$ is the observation matrix, and f is a random function uniformly distributed over the packing set. The quantity $I(y; f)$ is the mutual information between y and f, and a key step in the proofs is obtaining good upper bounds on it.

Let $\mathbb{P}_f$ (respectively $\mathbb{P}_g$) be the distribution of y given that $f^* = f$ (respectively $f^* = g$). The mutual information $I(y; f)$ is intimately related to the Kullback-Leibler (KL) divergence between $\mathbb{P}_f$ and $\mathbb{P}_g$, which is given by

$$D(\mathbb{P}_f \,\|\, \mathbb{P}_g) = \int p_f(y) \log \frac{p_f(y)}{p_g(y)} dy, \tag{63}$$

where p_f and p_g are the densities with respect to Lebesgue measure. Our analysis requires upper bounds on this KL divergence, as provided by the following lemma:

Lemma 14. *Assume that $\|f\|_\Phi = \|g\|_\Phi$. Then the Kullback-Leibler divergence is upper bounded as $D(\mathbb{P}_f \,\|\, \mathbb{P}_g) \leq \frac{n\|f-g\|_\Phi^2}{\sigma_m^2}$.*

See Appendix E.1 for the proof.

7.2 Proof of Theorem 3

We are now ready to begin the proof of our lower bounds on the minimax error in the (semi)-norm $\|\cdot\|_\Phi$. In order to leverage the lower bound (62), we need to have control on the packing and covering numbers in this norm:

Lemma 15 (Packing/covering in $\|\cdot\|_\Phi$-norm)**.** *Suppose that the kernel matrix K has polynomial-α decay.*

(a) Suppose that $m \leq (c_0 n)^{\frac{1}{2\alpha}}$ for some constant c_0. Then there exists a collection of functions $\{f^1, \ldots, f^M\}$ contained in $\mathbb{B}_\mathcal{H}(1)$ such that $M \geq 4^m$, and

$$\|f^i\|_\Phi^2 = \frac{\sigma_0^2}{16n} \quad \textit{and} \quad \|f^i - f^j\|_\Phi^2 \geq \frac{\sigma_0^2}{64n}, \quad \textit{for all } i \neq j \in \{1, 2, \ldots, M\}.$$

(b) The covering number of the set $\mathrm{Ra}(\Phi^*) \cap \mathbb{B}_\mathcal{H}(1)$ in the $\|\cdot\|_\Phi$-norm is upper bounded as

$$\log N_\Phi(\epsilon) \leq c_1 (1/\epsilon)^{\frac{1}{\alpha}}. \tag{64}$$

In the other direction, if $\epsilon^2 \geq \frac{\kappa_1}{m^{2\alpha}}$ for some constant $\kappa_1 > 0$, then the packing number is lower bounded as

$$\log M_\Phi(\epsilon) \geq c_2 (1/\epsilon)^{\frac{1}{\alpha}}. \tag{65}$$

The proof of this auxiliary result is given in Appendix E.2. We now use it to complete the proof of Theorem 3.

7.2.1 The case of time sampling

Let us consider part (a) first. Recall that in this case $\sigma_m = \sigma_0/\sqrt{m}$. First, supposing that $m \leq (c_0 n)^{\frac{1}{2\alpha}}$, we establish a lower bound of the order $1/n$ on the minimax risk. (Note that if this upper bound on m holds, then the $1/n$ term is the minimum of the two terms in Theorem 3(a).) Let $\{f^1, \ldots, f^M\}$ be the collection of functions from part Lemma 15(a). Using the Fano bound (62) and the inequality $\log M \geq m \log 4$, we obtain

$$\inf_{\widetilde{f}} \sup_{f^* \in \mathbb{B}_\mathcal{H}(1)} \mathbb{P}_{f^*}\Big[\|\widetilde{f} - f^*\|_\Phi^2 \geq \frac{\sigma_0^2}{256n}\Big] \geq 1 - \frac{I(y; f) + \log 2}{m \log 4},$$

where y is the matrix of observations $(y_1, \ldots, y_n) \in \mathbb{R}^{m \times n}$, and the random variable f ranges uniformly over the packing set $\{f^1, \ldots, f^M\}$. By the convexity of the Kullback-Leibler divergence, we have

$$I(y; f) \leq \frac{1}{\binom{M}{2}} \sum_{i \neq j} D(\mathbb{P}_{f^i} \,\|\, \mathbb{P}_{f^j}) \overset{(i)}{\leq} \frac{1}{\binom{M}{2}} \sum_{i \neq j} \frac{n\|f^i - f^j\|_\Phi^2}{\sigma_m^2} \overset{(ii)}{\leq} \frac{n}{\sigma_m^2} \frac{\sigma_0^2}{4n} = \frac{m}{4},$$

where inequality (i) follows from Lemma 14, and inequality (ii) follows from the packing construction in Lemma 15(a). Consequently, we have

$$\frac{I(y; f) + \log 2}{m \log 4} \leq \frac{m/4 + \log 2}{m \log 4} \leq \frac{1}{2}$$

for all $m \geq 2$, which completes the proof.

Otherwise, we may assume that $m \geq (c_0 n)^{\frac{1}{2\alpha}}$, under which assumption we prove the lower bound involving the term of order $(mn)^{-\frac{2\alpha}{2\alpha+1}}$. (Note that this lower bound on m holds, then the $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ term is the minimum of the two terms in Theorem 3(a).) Let $\delta^2 = c_3 (\frac{\sigma_0^2}{mn})^{\frac{2\alpha}{2\alpha+1}}$ for some $c_3 > 0$ to be chosen. Since $m \geq (c_0 n)^{\frac{1}{2\alpha}}$ by assumption, some algebra shows that $\delta^2 \geq \frac{\kappa_1}{m^{2\alpha}}$, so that the lower bound on the packing number from Lemma 15(b) may be applied. Combining this lower bound with the Fano inequality, we obtain

$$\inf_{\widetilde{f}} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}(1)} \mathbb{P}_{f^*} \big[\|\widetilde{f} - f^*\|_{\Phi}^2 \geq \frac{\delta^2}{4} \big] \geq 1 - \frac{I(y; f) + \log 2}{c_2 (1/\delta)^{1/\alpha}}.$$

By the upper bounding technique of Yang and Barron [33], the mutual information $I(y; f)$ is upper bounded by $\inf_{\nu > 0} \{\nu^2 + \log N_{\mathrm{KL}}(\nu)\}$, where N_{KL} is the covering number in the square-root Kullback-Leibler (pseudo)-metric. By Lemma 14 and Lemma 15(b), we have $N_{\mathrm{KL}}(\nu) \leq c_1 \big(\frac{\sigma_0}{\sqrt{nm}} \nu\big)^{1/\alpha}$. Re-parameterizing in terms of $\epsilon^2 = \frac{\sigma_0^2}{nm} \nu^2$, we obtain the upper bound

$$I(y; f) \leq \inf_{\epsilon > 0} \big\{ \frac{nm}{\sigma_0^2} \epsilon^2 + c_1 (1/\epsilon)^{1/\alpha} \big\} \;\leq\; \big(\frac{1}{\epsilon_*}\big)^{1/\alpha},$$

where $\epsilon_*^2 = c_4 \big(\frac{\sigma_0^2}{nm}\big)^{\frac{2\alpha}{2\alpha+1}}$ for some constant c_4. Consequently, we have

$$R := \frac{I(y; f) + \log 2}{c_2 (1/\delta)^{1/\alpha}} \leq \frac{\big(\frac{1}{\epsilon_*}\big)^{1/\alpha} + \log 2}{(1/\delta)^{1/\alpha}}.$$

Note that δ and ϵ_* are of the same order. By choosing the pre-factor c_3 sufficiently small, we can thus guarantee that the ratio R is less than 1/2, from which the claim follows.

7.2.2 The case of frequency truncation

Recall that in this case $\sigma_m = \sigma_0$. Since by assumption $m \geq (c_0 n)^{\frac{1}{2\alpha+1}}$, letting $\delta^2 = c_3 \big(\frac{\sigma_0^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}}$, we have $\delta^2 \geq \frac{\kappa_1}{m^{2\alpha}}$ after some algebra. Hence, the lower bound on the packing number from Lemma 15(b) may be applied. Moreover, we have $N_{\mathrm{KL}}(\nu) \leq c_1 \big(\frac{\sigma_0}{\sqrt{n}} \nu\big)^{1/\alpha}$. The rest of the proof follows that of part (a).

7.3 Proof of Theorem 4

On one hand, no method can estimate to an accuracy greater than $\frac{1}{2} N_m(\Phi)$. Indeed, whatever estimator $\widetilde{f}$ is used, the adversary can always choose some function f^* such that $\Phi(f^*) = 0$, and $\|\widetilde{f} - f^*\|_{L^2} \geq \frac{1}{2} N_m(\Phi)$. To see this, note that on one hand, if $\|\widetilde{f}\|_{L^2} \geq \frac{1}{2} N_m(\Phi)$, then the adversary can set $f^* = 0$. On the other hand, if $\|\widetilde{f}\|_{L^2} < \frac{1}{2} N_m(\Phi)$, then for any $\delta > 0$, adversary can choose a function $f^* \in \mathrm{Ker}(\Phi) \cap \mathbb{B}_{\mathcal{H}}(1)$ such that $\|f^*\|_{L^2} > N_m(\Phi) - \delta$, by definition (13) of $N_m(\Phi)$. We then have $\|f^* - \widetilde{f}\|_{L^2} \geq \|f^*\|_{L^2} - \|\widetilde{f}\|_{L^2} > \frac{1}{2} N_m(\Phi) - \delta$ where we let $\delta \to 0$. In addition, it follows from the theory of optimal widths in Hilbert spaces [26] that $N_m(\Phi) \gtrsim \mu_{m+1}$, thereby establishing the $m^{-2\alpha}$ lower bound for a kernel operator with polynomial-α decay.

Let us now prove the lower bound involving $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ in part (a). This term is the smaller of the two terms involved in the minimum, when $m \geq n^{\frac{1}{2\alpha}}$; this is the only case we need to consider as for $m < n^{\frac{1}{2\alpha}}$, the minimum is n^{-1} which is dominated by the term $m^{-2\alpha}$. We introduce the shorthand $\Psi_1^m = \mathrm{span}\{\psi_1, \dots, \psi_m\} \cap \mathbb{B}_{\mathcal{H}}(1)$, corresponding to the intersection of the unit ball $\mathbb{B}_{\mathcal{H}}(1)$ with the m-dimensional subspace of $\mathcal{H}$ spanned by the first m eigenfunctions of the kernel. For this proof, our packing/covering constructions take place entirely within this set. The following lemma, proved in Appendix E.3, provides bounds on these packing and covering numbers:

Lemma 16 (Packing/covering in $\|\cdot\|_{L^2}$-norm)**.** *There is a universal constant $c_1 > 0$ such that*

$$\log N_{L^2}(\epsilon; \Psi_1^m) \leq c_1 (1/\epsilon)^{\frac{1}{\alpha}}. \tag{66}$$

In the other direction, if $\epsilon^2 \geq \frac{\kappa_1}{m^{2\alpha}}$ for some constant $\kappa_1 > 0$, there is a universal constant $c_2 > 0$ such that

$$\log M_{L^2}(\epsilon; \Psi_1^m) \geq c_2 (1/\epsilon)^{\frac{1}{\alpha}}. \tag{67}$$

Based on this lemma, proving a $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ bound is relatively straightforward, once again using Fano's inequality (62). Choosing $\delta^2 = c_3(\frac{\sigma_0^2}{mn})^{\frac{2\alpha}{2\alpha+1}}$ for a constant c_3 to be specified, we construct a δ-packing in $\|\cdot\|_{L^2}$ norm, of size M such that $\log M \geq c_2(1/\delta)^{1/\alpha}$. As in the proof of Theorem 3, we upper bound the mutual information in terms of the covering number in the $\|\cdot\|_\Phi$. By condition (B2), this covering number is upper bounded (up to constant factors) by the covering number in the $\|\cdot\|_{L^2}$-norm. To see this, note that for any $f \in \Psi_1^m \cap \mathbb{B}_{L^2}(\varepsilon)$, we have $f = \sum_{j=1}^m a_j \psi_j$, with $\sum_{j=1}^m a_j^2/\mu_j \leq 1$ and $\sum_{j=1}^m a_j^2 \leq \varepsilon^2$. Then, condition (B2) implies $\|f\|_\Phi^2 = \langle a, \Psi a\rangle \leq c_1 \|a\|_2^2 \leq 2\varepsilon^2$, that is $f \in \Psi_1^m \cap \mathbb{B}_\Phi(\sqrt{c_1}\varepsilon)$. Finally, by Lemma 16, the $\|\cdot\|_{L^2}$ covering number scales as $(1/\epsilon)^{1/\alpha}$, so that the same calculations as before yield the $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ rate as claimed.

The proof of part (b) is similar. We only need to consider the case $m \geq n^{\frac{1}{2\alpha+1}}$. The rest of the argument follows by taking $\delta^2 = c_3\big(\frac{\sigma_0^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}}$ and recalling that $\sigma_m = \sigma_0$ in this case.

8 Discussion

We studied the problem of sampling for functional PCA from a functional-theoretic viewpoint. The principal components were assumed to lie in some Hilbert subspace $\mathcal{H}$ of L^2, usually a RKHS, and the sampling operator, a bounded linear map $\Phi : \mathcal{H} \to \mathbb{R}^m$. The observation model was taken to be the output of Φ plus some Gaussian noise. The two main examples of Φ considered were time sampling, $[\Phi f]_j = f(t_j)$, and (generalized) frequency truncation $[\Phi f]_j = \langle \psi_j, f\rangle_{L^2}$. We showed that it is possible to recover the subspace spanned by the original components, by applying a regularized version of PCA in $\mathbb{R}^m$ followed by simple linear mapping back to function space. The regularization involved the "trace-smoothness condition" (18) based on the matrix $K = \Phi\Phi^*$ whose eigendecay influenced the rate of convergence in $\mathbb{R}^m$.

We obtained the rates of convergence for the subspace estimators both in the discrete domain, $\mathbb{R}^m$, and the function domain, L^2. As examples, for the case of a RKHS $\mathcal{H}$ for which both the kernel integral operator and the kernel matrix K have polynomial-α eigendecay (i.e., $\mu_j \asymp \widehat{\mu}_j \asymp j^{-2\alpha}$), the following rates in HS-projection distance for subspaces in the function domain were worked out in details:

time sampling	frequency truncation
$\big(\frac{1}{mn}\big)^{\frac{2\alpha}{2\alpha+1}} + \big(\frac{1}{m}\big)^{2\alpha}$	$\big(\frac{1}{n}\big)^{\frac{2\alpha}{2\alpha+1}} + \big(\frac{1}{m}\big)^{2\alpha}$

The two terms in each rate can be associated, respectively, with the estimation error (due to noise) and approximation error (due to having finite samples of an infinite dimensional object). Both rates exhibit a trade-off between the number of statistical samples (n) and that of functional samples (m). The two rates are qualitatively different: the two terms in the time sampling case interact to give an overall fast rate of n^{-1} for the optimal trade-off $m \asymp n^{\frac{1}{2\alpha}}$, while there is no interaction between the two terms in the frequency truncation; the optimal trade-off gives an overall rate of $n^{-\frac{2\alpha}{2\alpha+1}}$, a characteristics of nonparametric problems. Finally, these rates were shown to be minimax optimal.

Acknowledgements

This research was partially supported by NSF grants CCF-0545862 and DMS–0907632 to MJW.

A A special kernel

In this appendix, we examine a simple reproducing kernel Hilbert space, corresponding to a Sobolev or spline class with smoothness $\alpha = 1$. We provide expressions for various approximation-theoretic quantities appearing in our results, such as $D_m(\Phi)$, $N_m(\Phi)$ and Ψ. Further background on the calculations given here can be found in the paper [2].

Let us consider the time sampling model (8) with uniformly spaced points $t_j = j/m$ for $j = 1, \ldots, m$. Elementary calculations show that $K = \big(m^{-1}\mathbb{K}(t_i, t_j)\big) = \frac{1}{m^2} LL^T$, where $L \in \mathbb{R}^{m \times m}$ is lower triangular with all the nonzero entries equal 1. It can be shown that the eigenvalues of K are given by $\widehat{\mu}_k := \big\{4m^2 \sin^2\big(\frac{\mu_k^{-1/2}}{2m+1}\big)\big\}^{-1}$ for $k = 1, 2, \ldots, m$. Using the inequalities $\frac{2}{\pi}x \leq \sin(x) \leq x$, for $0 \leq x \leq \pi/2$, we have

$$\Big(\frac{2m+1}{2m}\Big)^2 \mu_k \;\leq\; \widehat{\mu}_k \;\leq\; \frac{\pi^2}{4}\Big(\frac{2m+1}{2m}\Big)^2 \mu_k,$$

showing that $\widehat{\mu}_k$ is a good approximation of μ_k, even for moderate values of m.

Recalling the definition of $\Psi \in \mathbb{R}^{m \times m}$ from Section 4.2, it can be shown that it takes the form $\Psi = I_m + \frac{1}{m}\mathbb{I}_s \mathbb{I}_s^T$, where $\mathbb{I}_s \in \mathbb{R}^m$ is the vector with entries $[\mathbb{I}_s]_j = (-1)^{j+1}$. Since $\lambda_{\max}(\Psi) = 2$, condition (B2) is clearly satisfied.

Now we consider the quantity $D_m(\Phi)$; by Lemma 2, it suffices to bound the operator norm of $K^{-1/2}(K^2 - \Theta)K^{-1/2}$. Some algebra shows that $K^2 - \Theta = m^{-4}(\frac{1}{2}hh^T + \frac{1}{6}m^2 K)$, where $h = (1, 2, \ldots, m)$, so that

$$D_m(\Phi) = |\!|\!| K^{-1/2}(K^2 - \Theta)K^{-1/2} |\!|\!|_2 = \frac{1}{2m^4}\, h^T K^{-1} h + \frac{1}{6m^2} \;=\; \frac{1}{2m} + \frac{1}{6m^2} \;\leq\; \frac{1}{m}.$$

Finally, it can be shown that $N_m(\Phi) \precsim \frac{\log m}{m^2}$; see the paper [2] for details.

B Auxiliary lemmas

Here we collect the proofs of various auxiliary lemmas.

B.1 Proof of Lemma 1

The space $\mathrm{Ra}(\Phi^*)$ is finite-dimensional and hence closed, which guarantees validity of the well-known decomposition $\mathcal{H} = \mathrm{Ra}(\Phi^*) \oplus \mathrm{Ker}(\Phi)$. In particular, for any $f \in \mathcal{H}$, there is $a \in \mathbb{R}^m$ and $f^\perp \in \mathrm{Ker}(\Phi)$ such that $f = \Phi^* a + f^\perp$. Then, $\Phi f = Ka$, and

$$\|f\|_{\mathcal{H}}^2 \geq \|\Phi^* a\|_{\mathcal{H}}^2 = \langle \Phi^* a, \Phi^* a\rangle_{\mathcal{H}} = \langle a, \Phi\Phi^* a\rangle_{\mathbb{R}^m} = \langle Ka, Ka\rangle_K = \|\Phi f\|_K^2.$$

Equality holds iff $f^\perp = 0$ which gives the desired condition.

B.2 Proof of Lemma 2

By a well-known result, for a symmetric matrix, the numerical radius is equal to the operator norm. Thus, we have $|||K - K^{-1/2}\Theta K^{-1/2}|||_2 = \sup_{a\in\mathbb{R}^m\setminus\{0\}} \frac{|a^T\big(K-K^{-1/2}\Theta K^{-1/2}\big)a|}{\|a\|_2^2}$. Making the substitution $b = K^{-1/2}a$, or equivalently $a = K^{1/2}b$, we obtain

$$|||K - K^{-1/2}\Theta K^{-1/2}|||_2 = \sup_{b\in\mathbb{R}^m\setminus\{0\}} \frac{|b^T\big(K^2 - \Theta\big)b|}{b^T K b}$$

Now define the function $f = \Phi^* b \in \mathrm{Ra}(\Phi^*)$. With this definition, we have the following equivalences:

$$b^T K b = \|\Phi^* b\|_{\mathcal{H}}^2 \;=\; \|f\|_{\mathcal{H}}^2, \quad b^T K^2 b = \|\Phi f\|_2^2 \;=\; \|f\|_{\Phi}^2, \quad \text{and} \quad b^T\Theta b = \|\sum_{j=1}^m b_j\phi_j\|_{L^2}^2 \;=\; \|f\|_{L^2}^2,$$

from which the claim follows.

B.3 Proof of Lemma 3

The (truncated) QR decomposition [13] of Z^* has the form $Z^* = \widetilde{Z}^* R$, where $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$, and $R \in \mathbb{R}^{r\times r}$ is upper triangular with nonnegative diagonal entries. By construction, we have $\mathrm{Ra}(\widetilde{Z}^*) = \mathrm{Ra}(Z^*)$. Moreover, from the trace smoothness condition (18), we have

$$r\rho^2 \;\geq\; \sum_{j=1}^r \|z_j^*\|_K^2 \;=\; \mathrm{tr}\,\big((Z^*)^T K^{-1} Z^*\big) \;\geq\; \lambda_{\min}(R^T R)\,\mathrm{tr}\,\big((\widetilde{Z}^*)^T K^{-1}\widetilde{Z}^*\big) \tag{68}$$

where the final inequality follows from the bound (90) in Appendix I. Recalling the definition (25), we have $C_m(f^*) = \|(Z^*)^T Z^* - I_r\|_\infty = \|R^T R - I_r\|_\infty$. Since $\lambda_j(R^T R) = \lambda_j(R^T R - I_r) + 1$, we have

$$\max_{j=1,\ldots,r} |\lambda_j(R^T R) - 1| \leq |||R^T R - I_r|||_2 \leq r\|R^T R - I_r\|_\infty = rC_m(f^*). \tag{69}$$

Since $rC_m(f^*) \leq \frac{1}{2}$, we conclude that $\lambda_{\min}(R^T R) \geq \frac{1}{2}$. Combined with our earlier bound (68), we conclude that $\widetilde{Z}^*$ indeed satisfies the trace-smoothness condition.

B.4 Proof of Lemma 5

We only need to consider the case $\nu = 1$; the general case follows by rescaling. Consider the following local one-sided version of $D_m(\Phi)$,

$$U_{\mathrm{loc}}(\epsilon;\Phi) := \sup_{\substack{f \in \mathrm{Ra}(\Phi^*), \\ \|f\|_{\mathcal{H}} \le 1, \\ \|f\|_{\Phi}^2 \le \epsilon^2}} \|f\|_{L^2}^2. \tag{70}$$

Using an argument similar to that of Lemma 2, (70) is equivalent to

$$U_{\mathrm{loc}}(\epsilon;\Phi) = \sup_{\substack{b^T b \le 1, \\ b^T K b \le \epsilon^2}} b^T K^{-1/2} \Theta K^{-1/2} b. \tag{71}$$

Using Lagrange duality, we have

$$\begin{aligned} U_{\mathrm{loc}}(\epsilon;\Phi) &\le \inf_{t \ge 0} \big[\max\big(\lambda_{\max}(K^{-1/2}\Theta K^{-1/2} - tK), 0\big) + t\epsilon^2 \big] \\ &\le c_0 \epsilon^2 \end{aligned} \tag{72}$$

since (B1) implies $\lambda_{\max}(K^{-1/2}\Theta K^{-1/2} - c_0 K) \le 0$.

For $f \in \mathcal{H}$, let $f = g + f^{\perp}$ be its decomposition according to $\mathcal{H} = \mathrm{Ra}(\Phi^*) \oplus \mathrm{Ker}(\Phi)$. Then, $\|g\|_{\mathcal{H}}^2 + \|f^{\perp}\|_{\mathcal{H}}^2 = \|f\|_{\mathcal{H}}^2 \le 1$ and $\|f\|_{L^2}^2 \le 2\|g\|_{L^2}^2 + 2\|f^{\perp}\|_{L^2}^2$. Hence, we obtain

$$R_m(\epsilon;1) \le 2U_{\mathrm{loc}}(\varepsilon;\Phi) + 2N_m(\Phi). \tag{73}$$

Combining (72) and (73) proves the claim.

B.5 Proof of inequality (61)

By polarization identity and some algebra,

$$C_m(f^*) \le 2\rho^2 \sup_{\|f\|_{\mathcal{H}} \le 1,\ \|f\|_{L^2} = \frac{1}{\sqrt{2}\rho}} \big| \|f\|_{\Phi}^2 - \|f\|_{L^2}^2 \big|$$

Let $f = g + f^{\perp}$ be the decomposition according to $f \in \mathcal{H} = \mathrm{Ra}(\Phi^*) + \mathrm{Ker}(\Phi)$. Let $f \in \mathbb{B}_{\mathcal{H}}(1)$ and $\|f\|_{L^2} = \frac{1}{\sqrt{2}\rho}$. Then, as in Appendix B.4, we have $g, f^{\perp} \in \mathbb{B}_{\mathcal{H}}(1)$. Hence,

$$\big| \|f\|_{\Phi}^2 - \|f\|_{L^2}^2 \big| \le \underbrace{\big| \|g\|_{\Phi}^2 - \|g\|_{L^2}^2 \big|}_{\le D_m(\Phi)} + \big| \underbrace{\|g + f^{\perp}\|_{L^2}^2}_{a^2} - \underbrace{\|g\|_{L^2}^2}_{b^2} \big|$$

where we have define $a, b > 0$ as above for simplicity. Let $d := \|f^{\perp}\|_{L^2}$. By triangle inequality, $b \le a + d$ and $|a - b| \le d$. Then,

$$|a^2 - b^2| = |a - b|(a + b) \le d(2a + d) \le \Big(\sqrt{\frac{2}{\rho}} + 1\Big) N_m(\Phi),$$

since $a = \frac{1}{\sqrt{2}\rho}$ and $d \le N_m(\Phi) \le 1$, by assumption.

C Proofs for Theorem 1

In this appendix, we collect the proofs of various auxiliary lemmas involved in the proof of Theorem 1.

C.1 Derivation of the bound (51)

From the CS-decomposition (44), we have $\widehat{Z}^T \widetilde{Z}^* = \widehat{C}$, and hence $\widehat{P}\widetilde{Z}^* = \widehat{Z}\widehat{C} - \widetilde{Z}^* = \widehat{E}\widehat{C} - \widetilde{Z}^*(I_r - \widehat{C})$. From the decomposition (43), we have

$$\begin{aligned}\langle\!\langle \widehat{P},\ \Delta_1 \rangle\!\rangle &= \sigma \operatorname{tr}\big[\overline{W} S R^T (\widetilde{Z}^*)^T \widehat{P} + \widetilde{Z}^* R S \overline{W}^T \widehat{P}\big] \\ &= 2\sigma \operatorname{tr}\big[R S \overline{W}^T \widehat{P} \widetilde{Z}^*\big] \\ &= 2\sigma \Big\{ \operatorname{tr}\big[R S \overline{W}^T \widehat{E}\widehat{C}\big] - \operatorname{tr}\big[R S \overline{W}^T \widetilde{Z}^* (I_r - \widehat{C})\big] \Big\},\end{aligned}$$

where we have used the standard facts $\operatorname{tr}(AB^T) = \operatorname{tr}(A^T B)$ and $\operatorname{tr}(AB) = \operatorname{tr}(BA)$. For the first term we have

$$\Big| \operatorname{tr}\big[R S \overline{W}^T \widehat{E}\widehat{C}\big] \Big| = \Big| \sum_{j,k=1}^{r} R_{jk} s_k \langle \overline{w}_k, \widehat{e}_j \rangle \, \widehat{c}_j \Big| \leq \Big(\sum_{j,k=1}^{r} R_{j,k}^2 \Big)^{1/2} \Big(\sum_{j,k} s_k^2 \, \widehat{c}_j^2 \, (\langle \overline{w}_k, \widehat{e}_j \rangle)^2 \Big)^{1/2}$$

where we have used Cauchy-Schwarz. By (69), under the assumption $r C_m(f^*) \leq \frac{1}{2}$, we have $\operatorname{tr}(R^T R) \leq \frac{3}{2} r$. We also have $0 < s_k \leq s_1 = 1$ and $0 \leq \widehat{c}_j \leq 1$ for $j, k = 1, \ldots, r$. It follows that

$$\Big| \operatorname{tr}\big[R S \overline{W}^T \widehat{E}\widehat{C}\big] \Big| \leq \sqrt{\frac{3}{2}} \sqrt{r} \Big(\sum_{j,k=1}^{r} (\langle \overline{w}_k, \widehat{e}_j \rangle)^2 \Big)^{1/2} \leq \sqrt{\frac{3}{2}} \, r^{3/2} \max_{j,k} |\langle \overline{w}_k, \widehat{e}_j \rangle|.$$

For the second term, using a similar argument by applying Cauchy-Schwarz, we get

$$\begin{aligned}\Big| \operatorname{tr}\big[R S \overline{W}^T \widetilde{Z}^* (I_r - \widehat{C})\big] \Big| &\leq \sqrt{\frac{3}{2}} \sqrt{r} \Big(\sum_{j=1}^{r} (1 - \widehat{c}_j)^2 \sum_{k=1}^{r} (\langle \overline{w}_k, \widetilde{z}_j^* \rangle)^2 \Big)^{1/2} \\ &\leq \sqrt{\frac{3}{2}} \, r \Big(\sum_{j} (1 - \widehat{c}_j)^2 \Big)^{1/2} \max_{j,k} |\langle \overline{w}_k, \widetilde{z}_j^* \rangle| \leq \frac{\sqrt{3}}{2\sqrt{2}} \, r \, |\!|\!| \widehat{E} |\!|\!|_{HS}^2 \max_{j,k} |\langle \overline{w}_j, \widetilde{z}_k^* \rangle|.\end{aligned}$$

where the last inequality follows from the fact that $\big(\sum_j (1 - \widehat{c}_j)^2\big)^{1/2} \leq \sum_j (1 - \widehat{c}_j) = \frac{1}{2} |\!|\!| \widehat{E} |\!|\!|_{HS}^2$.

C.2 Proof of Lemma 8

We make use of an ellipsoid approximation (see [23]). To simplify notation, define $\widetilde{K} := (8 r \rho^2) K$ and $\widetilde{\mu} := 8 r \rho^2 \, \widehat{\mu}$, so that we have $\operatorname{tr}(Z^T K^{-1} Z) \leq 2 r \rho^2$ if and only if $\operatorname{tr}(Z^T \widetilde{K}^{-1} Z) \leq 1/4$. Since both $\widehat{Z}$ and $\widetilde{Z}^*$ satisfy this condition, it follows that $\|\overline{z}_j\|_{\widetilde{K}} \leq \frac{1}{2}$ and $\|\widetilde{z}_j^*\|_{\widetilde{K}} \leq \frac{1}{2}$ for $j = 1, \ldots, r$, where $\|a\|_{\widetilde{K}}^2 := a^T \widetilde{K}^{-1} a$. Thus, we are guaranteed that $\widehat{e}_j \in \mathcal{E}_{\widetilde{K}} := \big\{ v \in \mathbb{R}^m \mid \|v\|_{\widetilde{K}} \leq 1 \big\}$.

We first establish an upper bound on the quantity $\sup \big\{ \langle \overline{w}_k, v \rangle \ \mid \ v \in \mathcal{E}_{\widetilde{K}} \cap \mathbb{B}_2(t) \big\}$, where $\mathbb{B}_2(t) = \{ v \in \mathbb{R}^m \mid \|v\|_2 \leq t \}$ is the Euclidean ball of radius t. Let $\widetilde{\mu}_1 \geq \cdots \geq \widetilde{\mu}_m$ be the eigenvalues of $\widetilde{K}$ in decreasing order and let $\widetilde{\mu} := (\widetilde{\mu}_1, \ldots, \widetilde{\mu}_m)$. Since for $U \in V_m(\mathbb{R}^m)$, the random vectors $\overline{w}_k$ and $U \overline{w}_k$ have the same distribution, it is equivalent to bound the quantity

$\sup\big\{\langle \overline{w}_k, v\rangle \;\mid\; v \in \mathcal{E}_{\widetilde{\mu}} \cap \mathbb{B}_2(t)\big\}$. Now for $v \in \mathcal{E}_{\widetilde{\mu}} \cap \mathbb{B}_2(t)$, we have $\sum_{i=1}^m \widetilde{\mu}_i^{-1} v_i^2 \le 1$ and $\sum_{i=1}^m t^{-2} v_i^2 \le 1$ implying $\sum_{i=1}^m \max\{\widetilde{\mu}_i^{-1}, t^{-2}\} v_i^2 \le 2$. Consequently, if we define the modified ellipse $\mathcal{E}_\gamma := \big\{v \in \mathbb{R}^m \mid \sum_{i=1}^m \frac{v_i^2}{\gamma_i} \le 1\big\}$ where $\gamma_i := 2\min\{t^2, \widetilde{\mu}_i\}$, then we are guaranteed that $v \in \mathcal{E}_\gamma$, so that it suffices to upper bound $\sup_{v \in \mathcal{E}_\gamma} \langle \overline{w}_k, v\rangle$. For future reference, we note that

$$\|\gamma\|_1 = 16\,\mathcal{F}^2(t/\sqrt{8}), \quad \text{and} \quad \|\gamma\|_\infty \le 2t^2 \tag{74}$$

where $\mathcal{F}$ was defined previously (24). Define the random variables $\overline{w}_k := \frac{1}{n}\sum_{i=1}^n w_i \beta_{ik}$ and $\overline{B}_{kk} := \frac{1}{n}\sum_{i=1}^n \beta_{ik}^2$. For each index k, Lemma 17 (see Appendix F), combined with the relations (74), yields

$$\sigma \sup_{v \in \mathcal{E}_\gamma} |\langle \overline{w}_k, v\rangle| \le \sigma \overline{B}_{kk}^{1/2}\Big\{\sqrt{\frac{\|\gamma\|_1}{n}} + \delta\sqrt{\frac{\|\gamma\|_\infty}{n}}\Big\} \;\le\; C_1\, \overline{B}_{kk}^{1/2} \frac{\sigma}{\sqrt{n}}\big\{\mathcal{F}(t/\sqrt{8}) + \delta t\big\}$$

with probability at least $1 - \exp(-\delta^2/2)$. Taking $\delta = A_r \sqrt{n}\, t/\sigma$, where $A_r := \kappa r^{-3/2}$ for some small enough constant $\kappa > 0$, we obtain

$$\sigma \sup_{v \in \mathcal{E}_\gamma} |\langle \overline{w}_k, v\rangle| \le C_1\, \overline{B}_{kk}^{1/2}\Big\{\frac{\sigma}{\sqrt{n}}\mathcal{F}(t/\sqrt{8}) + A_r t^2\Big\}$$

with probability at least $1 - \exp(-A_r^2\, n\, t^2/2\sigma^2)$.

As was mentioned earlier, the same bound with the same probability holds for $\sup\big\{\sigma|\langle \overline{w}_k, v\rangle \mid v \in \mathcal{E}_{\widetilde{K}} \cap B_2(t)\big\}$. Since $\widehat{e}_j \in \mathcal{E}_{\widetilde{K}}$, $j = 1, \dots, r$ we can apply the technical Lemma 20 of Appendix H with $\nu = (n, m)$, $\theta_\nu = A_r n/\sigma^2$ and $t_\nu = \epsilon_{m,n}$ to obtain

$$\sigma|\langle \overline{w}_k, \widehat{e}_j\rangle| \le C_1\, \overline{B}_{kk}^{1/2}\Big\{\frac{\sigma}{\sqrt{n}}\mathcal{F}(2\|\widehat{e}_j\|_2/\sqrt{8}) + A_r(2\|\widehat{e}_j\|_2)^2 + \frac{\sigma}{\sqrt{n}}\mathcal{F}(2\epsilon_{m,n}/\sqrt{8}) + A_r(2\epsilon_{m,n})^2\Big\},$$

for all $j \in \{1, \dots, r\}$, with probability at least $1 - c_1 \exp(-A_r^2\, n\, \epsilon_{m,n}^2/2\sigma^2)$. Note that $\|\widehat{e}_j\|_2 \le |\!|\!|\widehat{E}|\!|\!|_{HS}$, $j = 1, \dots, r$. Since the bound obtained above is nondecreasing in $\|\widehat{e}_j\|$, we can replace $\|\widehat{e}_j\|$ everywhere with $|\!|\!|\widehat{E}|\!|\!|_{HS}$. We also note that by χ_n^2 concentration [17, 19], we have $\overline{B}_{kk} \le 3/2$ with probability at least $1 - \exp(-n/64)$. Finally, by definition of $\epsilon_{m,n}$ and monotonicity of $\mathcal{F}$ we have $\frac{\sigma}{\sqrt{n}}\mathcal{F}(2\epsilon_{m,n}/\sqrt{8}) \le \frac{\sigma}{\sqrt{n}}\mathcal{F}(\epsilon_{m,n}) \le A_r \epsilon_{m,n}^2$. Putting together the pieces, we conclude that

$$\max_{j,k} \sigma|\langle \overline{w}_k, \widehat{e}_j\rangle| \le C_2\Big\{\frac{\sigma}{\sqrt{n}}\mathcal{F}(|\!|\!|\widehat{E}|\!|\!|_{HS}) + A_r |\!|\!|\widehat{E}|\!|\!|_{HS}^2 + A_r\, \epsilon_{m,n}^2\Big\},$$

with probability at least $1 - c_1 r \exp(-A_r^2 n\, \epsilon_{m,n}^2/2\sigma^2) - r\exp(-n/64)$, where we have used union bound to obtain a uniform result over k.

C.3 Proof of Lemma 9

We control terms of the form $\langle \overline{w}_k, \widetilde{z}_j^*\rangle$ using Lemma 17 in Appendix F, this time with w_i replaced with $\langle w_i, \widetilde{z}_j^*\rangle$ and $\gamma = 1$ (i.e., we are looking at sums of products of univariate Gaussians). Thus, for any fixed j and k, we have $\sigma|\langle \overline{w}_k, \widetilde{z}_j^*\rangle| \le \sigma\, \overline{B}_{kk}^{1/2}\{\frac{1}{\sqrt{n}} + \delta\frac{1}{\sqrt{n}}\}$, with probability at least $1 - \exp(-\delta^2/2)$. Taking $\delta = \kappa r^{-1}\sqrt{n}/\sigma$, then the event $\max_k \overline{B}_{kk} \le 3/2$, which we have already accounted for, we have by union bound

$$\max_{j,k} \sigma r|\langle \overline{w}_k, \widetilde{z}_j^*\rangle| \le \sqrt{\frac{3}{2}}\Big\{\frac{\sigma}{\sqrt{n}} r + \kappa\Big\} \le \sqrt{6}\kappa$$

with probability at least $1 - r^2\exp(-\kappa^2 r^{-2} n/2\sigma^2)$. The second inequality follows by our assumption $r \le \kappa\sqrt{n}/\sigma$.

C.4 Proof of Lemma 10

For each $j \in \{1, \dots, r\}$, we define the vector $\zeta^j := W\widetilde{z}_j^* \in \mathbb{R}^n$ so that $\zeta^j = (\zeta_i^j)$ where $\zeta_i^j = w_i^T \widetilde{z}_j^*$. We can use the same ellipsoid approximation as Appendix C.2— that is, we first look at $\sup\big\{T_1(v; \widetilde{z}_j^*) \;\mid\; v \in \mathcal{E}_{\widetilde{K}} \cap \mathbb{B}_2(t)\big\}$ and then argue that it is enough to bound $\sup_{v\in\mathcal{E}_\gamma} T_1(v; \widetilde{z}_j^*) = \sup_{v\in\mathcal{E}_\gamma} \langle v, \frac{1}{n}\sum_{i=1}^n \zeta_i^j w_i - \widetilde{z}_j^*\rangle$, due to the invariance of the underlying distribution under orthogonal transformations of v. Now applying Lemma 18 from Appendix F yields

$$\sigma^2 \sup_{v\,\in\,\mathcal{E}_\gamma} T_1(v; \widetilde{z}_j^*) \;\le\; \Big(\frac{\|\zeta^j\|_2}{\sqrt{n}} + 1\Big)\Big\{\frac{\sigma^2}{\sqrt{n}}\sqrt{\|\gamma\|_1} + \delta\sigma^2\sqrt{\|\gamma\|_\infty}\Big\} \tag{75}$$

with probability at least $1 - 2\exp(-n\frac{\delta\wedge\delta^2}{16})$. Recalling that by assumption $\sigma \le \sigma_0$, let $\widetilde{A}_r = \kappa r^{-1}$. For $t \le \sigma\sigma_0/\widetilde{A}_r$, take $\delta = \widetilde{A}_r t/(\sigma\sigma_0) \le 1$. Then, using (74), the left-hand side of (75) is bounded above by

$$C_1\Big(\frac{\|\zeta^j\|_2}{\sqrt{n}} + 1\Big)\Big\{\sigma_0\frac{\sigma}{\sqrt{n}}\mathcal{F}(t/\sqrt{8}) + \widetilde{A}_r t^2\Big\} \tag{76}$$

with probability at least $1 - 2\exp\big(-\widetilde{A}_r^2 n\, t^2/(16\,\sigma^2\sigma_0^2)\big)$. For $t > \sigma\sigma_0/\widetilde{A}_r$, take $\delta = \widetilde{A}_r t/\sigma^2$. In this case, $\widetilde{A}_r t > \sigma\sigma_0 \ge \sigma^2$ implying $\delta > 1$. Then, the left-hand side of (75) is again bounded above by (76), this time with probability at least $1 - 2\exp(-\widetilde{A}_r n\, t/(16\,\sigma^2))$. Assuming $\kappa \le 1$, which is going to be the case, we have $\widetilde{A}_r^2 \le \widetilde{A}_r$. Combining the two cases, we have the upper bound (76) with probability at least

$$1 - 2\underbrace{\exp\big\{-\widetilde{A}_r^2\, n\,(\sigma_0^{-2}\wedge 1)\,(t\wedge t^2)\,/\,(16\sigma^2)\big\}}_{p_1(t)} \tag{77}$$

for all $t > 0$. (Note the break-up into two cases was to obtain a dependence of σ^{-2} in the probability exponent for all $t > 0$.)

By an argument similar to Appendix C.2—that is, using technical Lemma 20—we have $\|\widehat{e}_j\|_2 \le \|\!|\widehat{E}|\!\|_{HS}$ and $\frac{\sigma}{\sqrt{n}}\mathcal{F}(2\epsilon_{m,n}/\sqrt{8}) \le \widetilde{A}_r\epsilon_{m,n}^2$ from the definition; we obtain

$$\sigma^2 T_1(\widehat{e}_j; \widetilde{z}_j^*) \le C_2\big(\frac{\|\zeta^j\|_2}{\sqrt{n}} + 1\big)\big\{\sigma_0\frac{\sigma}{\sqrt{n}}\mathcal{F}(\|\!|\widehat{E}|\!\|_{HS}) + \widetilde{A}_r\|\!|\widehat{E}|\!\|_{HS}^2 + \widetilde{A}_r\epsilon_{m,n}^2\big\}$$

for all $j \in \{1, \dots, r\}$, with probability at least that of (77) with $t = \epsilon_{m,n}$ and 2 replaced with some constant $c_2 > 2$, i.e. $1 - c_2\, p_1(\epsilon_{m,n})$. By concentration of χ_n^2 variables and union bound, we have $\max_j n^{-1}\|\zeta^j\|_2^2 \le 3/2$ with probability at least $1 - r\exp(-n/64)$. Putting together the pieces, we conclude that

$$\sigma^2\sum_{j=1}^r T_1(\widehat{e}_j; \widetilde{z}_j^*) \le C_3\Big\{\sigma_0\frac{\sigma}{\sqrt{n}} r\mathcal{F}(\|\!|\widehat{E}|\!\|_{HS}) + \kappa\|\!|\widehat{E}|\!\|_{HS}^2 + \kappa\epsilon_{m,n}^2\Big\}$$

with probability at least $1 - c_2\, p_1(\epsilon_{m,n}) - r\exp(-n/64)$, as claimed.

C.5 Proof of Lemma 11

As before, the problem of bounding $T_2(\widehat{e}_j)$ can be reduced to controlling $\sup_{v \in \mathcal{E}_\gamma} T_2(v)$, by invariance under orthogonal transformation. Applying Lemma 19 of Appendix G with with $\delta = \kappa\sqrt{n}\, t/\sigma$ yields

$$\begin{aligned}\sigma \sup_{v \in \mathcal{E}_\gamma} \sqrt{T_2(v)} = \sup_{v\,\in\,\mathcal{E}_\gamma} \frac{\sigma}{\sqrt{n}} \|Wv\|_2 &\leq \sigma \Big\{ \sqrt{\frac{\|\gamma\|_1}{n}} + \big(1 + \kappa \frac{t}{\sigma}\big) \sqrt{\|\gamma\|_\infty} \Big\} \\ &\leq C_1 \big\{ \frac{\sigma}{\sqrt{n}} \mathcal{F}(t/\sqrt{8}) + \sigma t + \kappa t^2 \big\} \\ &\leq C_1 \big\{ \frac{\sigma}{\sqrt{n}} \mathcal{F}(t) + \sigma t + \sqrt{2}\kappa t \big\}\end{aligned}$$

with probability at least $1 - \exp(-\kappa^2 n\, t^2/2\sigma^2)$, valid for all $t \leq \sqrt{2}$, Note that since $\|\widehat{e}_j\|_2 \leq \sqrt{2}$ (by proper alignment), it is enough to only have a bound for $t \leq \sqrt{2}$. Recall the assumption (A1), $\sigma \leq \sqrt{\kappa}$ and by (A3), $\frac{\sigma}{\sqrt{n}} \mathcal{F}(t) \leq \sqrt{\kappa} t$. Assuming $\kappa < 1$, we obtain

$$\sigma^2 \sup_{v\,\in\,\mathcal{E}_\gamma} T_2(v) \leq C_1^2 \big(2\sqrt{\kappa}\, t + \sqrt{2}\kappa t\big)^2 \leq C_2 \kappa t^2$$

with the same probability. As before, applying technical Lemma 20, this time with $t_\nu = r^{-1/2}\epsilon_{m,n}$, we obtain

$$\sigma^2 T_2(\widehat{e}_j) \;\leq\; C_2 \kappa \Big\{ (2\|\widehat{e}_j\|_2)^2 + \Big(2 \frac{\epsilon_{m,n}}{\sqrt{r}}\Big)^2 \Big\}$$

for all $j \in \{1, \ldots, m\}$ with probability at least $1 - c_3 \exp(-\kappa^2 r^{-2} n \epsilon_{m,n}^2 / 2\sigma^2)$. Thus, we have

$$\sigma^2 \sum_{j=1}^{r} T_2(\widehat{e}_j) \leq C_3 \kappa \Big\{ \|\!|\widehat{E}|\!\|_{HS}^2 + \epsilon_{m,n}^2 \Big\}$$

with probability the same probability. Note that we have used $\|\!|\widehat{E}|\!\|_{HS}^2 = \sum_j \|\widehat{e}_j\|_2^2$.

D Proofs for Theorem 2

In this appendix, we collect the proofs of various auxiliary lemmas involved in the proof of Theorem 2.

D.1 Proof of Lemma 12

By definition, each $\widehat{h}_j$ lies in $\widehat{\mathfrak{F}}$, so that we have $\widehat{h}_j = \Phi^*\big(\sum_i B_{ij} K^{-1} \widehat{z}_i\big)$ for some $B \in \mathbb{R}^{r \times r}$. Recalling that $[K^{-1}\widehat{Z}B]_j$ denotes the j-th column of $K^{-1}\widehat{Z}B$, we can write $\widehat{h}_j = \Phi^*[K^{-1}\widehat{Z}B]_j$. Recalling the formula (5) for the adjoint, observe that for any $a, b \in \mathbb{R}^m$, we have

$$\langle \Phi^* a, \Phi^* b \rangle_{L^2} = \langle \textstyle\sum_i a_i \varphi_i, \sum_j b_j \varphi_j \rangle_{L^2} = \sum_{i,j} a_i b_j \langle \varphi_i, \varphi_j \rangle_{L^2} = a^T \Theta\, b \qquad (78)$$

where $\Theta = (\langle \varphi_i, \varphi_j \rangle_{L^2}) \in \mathbb{S}_+^m$, as previously defined in Lemma 2. Since the functions $\{\widehat{h}_j\}_{j=1}^r$ are orthonormal in L^2, we must have $\langle \widehat{h}_j, \widehat{h}_k \rangle_{L^2} = [K^{-1}\widehat{Z}B]_j^T \Theta [K^{-1}\widehat{Z}B]_k \;=\; \delta_{jk}$, or in matrix form $(K^{-1}\widehat{Z}B)^T \Theta (K^{-1}\widehat{Z}B) = I_{r \times r}$. This condition can be re-written as

$$B^T \widehat{Q} B = I_{r \times r} = I, \quad \text{where} \quad \widehat{Q} := \widehat{Z}^T K^{-1} \Theta K^{-1} \widehat{Z}.$$

Since $\widehat{h}_j - \widehat{g}_j = \Phi^*[K^{-1}\widehat{Z}(B - I_r)]_j$, we have $\|\widehat{h}_j - \widehat{g}_j\|_{L^2}^2 = [K^{-1}\widehat{Z}(B - I_r)]_j^T \Theta [K^{-1}\widehat{Z}(B - I_r)]_j$, using the definition of Θ. Consequently, we obtain

$$\sum_{j=1}^r \|\widehat{h}_j - \widehat{g}_j\|_{L^2}^2 = \operatorname{tr}\left\{\big(K^{-1}\widehat{Z}(B-I)\big)^T \Theta \big(K^{-1}\widehat{Z}(B-I)\big)\right\} = \operatorname{tr}\left\{I + \widehat{Q} - 2\widehat{Q}B\right\},$$

using the symmetry of $\widehat{Q}$ and the constraint $B^T\widehat{Q}B = I$. Subject to this constraint, we are free to choose B as we please; setting $B = \widehat{Q}^{-1/2}$ yields

$$\sum_{j=1}^r \|\widehat{h}_j - \widehat{g}_j\|_{L^2}^2 = \operatorname{tr}\left\{\big(I - \widehat{Q}^{1/2}\big)^2\right\} = |||I - \widehat{Q}^{1/2}|||_{HS}^2.$$

In order to upper bound $|||I - \widehat{Q}^{1/2}|||_{HS}$, we first control the closely related quantity $|||I - \widehat{Q}|||_{HS}$. We have

$$\begin{aligned} |||I_r - \widehat{Q}|||_{HS} &= |||\widehat{Z}^T K^{-1/2}(K - K^{-1/2}\Theta K^{-1/2})K^{-1/2}\widehat{Z}|||_{HS} \\ &\leq |||K - K^{-1/2}\Theta K^{-1/2}|||_2 \, |||\widehat{Z}^T K^{-1}\widehat{Z}|||_{HS} \\ &\leq 2r\rho^2 \, D_m(\Phi), \end{aligned} \tag{79}$$

where we have used inequality (93), Lemma 2, the trace-smoothness condition $\operatorname{tr}(\widehat{Z}^T K^{-1}\widehat{Z}) \leq 2r\rho^2$, and the inequality $|||M|||_{HS} \leq \operatorname{tr}(M)$, valid for any $M \succeq 0$.

In order to bound $|||I - \widehat{Q}^{1/2}|||_{HS}$, we apply the inequality

$$|||A^q - D^q||| \leq q a^{q-1} |||A - D|||, \quad 0 < q < 1, \tag{80}$$

valid for any operators A, D such that $A \succeq aI$ and $D \succeq aI$ for some positive number a, where $|||\cdot|||$ is any unitarily invariant norm. (See Bhatia [5], equation (X.46) on p. 305). As long as $2r\rho^2 D_m(\Phi) \leq 1/2$ so that the bound (79) implies that $\widehat{Q} \succeq 1/2I$, we may apply the inequality (80) with $A = I_r$, $D = \widehat{Q}$, $a = q = 1/2$ and $|||\cdot||| = |||\cdot|||_{HS}$ so as to obtain the inequality $|||I_r - \widehat{Q}^{1/2}|||_{HS} \leq \frac{1}{\sqrt{2}}|||I_r - \widehat{Q}|||_{HS}$, which completes the proof.

D.2 Proof of Lemma 13

By definition, we have $h_j^* = \sum_{i=1}^r E_{ij} f_i^*$ for some $E \in \mathbb{R}^{r \times r}$. Since both $\{h_j^*\}$ and $\{f_j^*\}$ are assumed orthonormal in L^2, the matrix E must be orthonormal. In addition, we have $\sum_{j=1}^r \|h_j^*\|_{\mathcal{H}}^2 = \sum_{j=1}^r \|f_j^*\|_{\mathcal{H}}^2 \leq r\rho^2$, implying that

$$\sum_{j=1}^r \|h_j^* - g_j^*\|_{\mathcal{H}}^2 \leq 2\sum_{j=1}^r \big(\|h_j^*\|_{\mathcal{H}}^2 + \|g_j^*\|_{\mathcal{H}}^2\big) \leq 2(r\rho^2 + 2r\rho^2) \leq 6r\rho^2$$

where we have used the fact that $\sum_{j=1}^r \|g_j^*\|_{\mathcal{H}}^2 = \sum_{j=1}^r \|\widetilde{z}_j^*\|_K^2 \leq 2r\rho^2$.

Recall the argument leading to the bound (56); applying this same reasoning to the pair (h_j^*, g_j^*) with the choices $A^2 = \sum_{j=1}^r \|h_j^* - g_j^*\|_\Phi^2$ and $B^2 = 6r\rho^2$ leads to

$$\sum_{j=1}^r \|h_j^* - g_j^*\|_{L^2}^2 \leq c_1 \sum_{j=1}^r \|h_j^* - g_j^*\|_\Phi^2 + 6r\rho^2 S_m(\Phi).$$

It remains to bound the term $\sum_{j=1}^{r} \|h_j^* - g_j^*\|_\Phi^2$. Recalling that $\Phi f_i^* = z_i^*$, we note that $\Phi h_j^* = \sum_i E_{ij} z_i^* = [Z^* E]_j$. It follows that $\sum_{j=1}^{r} \|h_j^* - g_j^*\|_\Phi^2 = \|Z^* E - \widetilde{Z}^*\|_{HS}^2$. Since $\mathrm{Ra}(Z^*) = \mathrm{Ra}(\widetilde{Z}^*)$, there exists a matrix $R \in \mathbb{R}^{r \times r}$ such that $Z^* = \widetilde{Z}^* R$. Letting $V_1 \Upsilon V_2^T$ denote the SVD of R, we have

$$\|Z^* E - \widetilde{Z}^*\|_{HS} = \|RE - I_r\|_{HS} = \|R - E^T\|_{HS} = \|V_1 \Upsilon V_2^T - E^T\|_{HS},$$

where we have used the unitary invariance of the Hilbert-Schmidt norm. Take $E^T = V_1 V_2^T$ which is orthogonal, hence a valid choice. By unitary invariance, we have $\|Z^* E - \widetilde{Z}^*\|_{HS} = \|\Upsilon - I_r\|_{HS}$. We now apply inequality (80) with $a = q = 1/2$, $A = \Upsilon^2$, $D = I_r$, $|\!|\!| \cdot |\!|\!| = |\!|\!| \cdot |\!|\!|_{HS}$. The condition $r C_m(f^*) \leq \frac{1}{2}$ implies $\Upsilon^2 \succeq \frac{1}{2} I_r$. (See Appendix B.3, in particular the argument following (69).) Consequently, we have $\|\Upsilon - I_r\|_{HS} \leq \frac{1}{\sqrt{2}} \|\Upsilon^2 - I_r\|_{HS} \leq \frac{1}{\sqrt{2}} |\!|\!| Z^{*T} Z^* - I_r |\!|\!|_{HS}$, where we have used $V_2 \Upsilon^2 V_2^T = R^T R = Z^{*T} Z^*$. Recalling that $|\!|\!| Z^{*T} Z^* - I_r |\!|\!|_{HS} \leq r C_m(f^*)$ and putting together the pieces, we obtain the stated inequality (59).

E Proofs for Theorems 3 and 4

In this appendix, we prove various lemmas that are involved in the proofs of the lower bounds given in Theorems 3 and 4.

E.1 Proof of Lemma 14

Let us introduce the shorthand notation $u = \Phi(f)$ and $v = \Phi(g)$. Under the model $\mathbb{P}_f$, for each $i = 1, 2, \ldots, n$, the vector $y_i \in \mathbb{R}^m$ has a zero-mean Gaussian distribution with covariance matrix $\Sigma_f := uu^T + \sigma_m^2 I$. Similarly, under the model $\mathbb{P}_g$, it is zero-mean Gaussian with covariance $\Sigma_g := vv^T + \sigma_m^2 I$. Since the data is i.i.d. and using standard formula for the Kullback-Leibler divergence between multivariate Gaussian distributions, we have $\frac{2}{n} D(\mathbb{P}_f \,\|\, \mathbb{P}_g) = \log \frac{\det \Sigma_g}{\det \Sigma_f} + \mathrm{tr}(\Sigma_g^{-1} \Sigma_f) - m$. Since $\|u\|_2 = \|v\|_2$ by construction, the matrices Σ_f and Σ_g have the same eigenvalues, and so the first term vanishes. Using the matrix inversion formula, we have

$$\frac{2}{n} D(\mathbb{P}_f \,\|\, \mathbb{P}_g) + m \;=\; \langle\!\langle (\sigma_m^2 I + vv^T)^{-1},\, \sigma_m^2 I + uu^T \rangle\!\rangle = \langle\!\langle \sigma_m^{-2} I - \sigma_m^{-4} \frac{vv^T}{1 + \|v\|_2^2 \sigma_m^{-2}},\, \sigma_m^2 I + uu^T \rangle\!\rangle$$

and some algebra, using the fact that $\|u\|_2 = \|v\|_2 = a$ implies $|\langle u, v \rangle| \leq a^2$, yields the claim.

E.2 Proof of Lemma 15

As previously observed, any function $f \in \mathrm{Ra}(\Phi^*) \cap \mathbb{B}_{\mathcal{H}}(1)$ can be represented by a vector in the ellipse $\mathcal{E} := \{\theta \in \mathbb{R}^m \mid \sum_{j=1}^{m} \theta_j^2 / \widehat{\mu}_j \leq 1\}$ such that $\|f\|_\Phi = \|\theta\|_2$. The proofs of both parts (a) and (b) exploit this representation.

(a) Note that the ellipse $\mathcal{E}$ contains the ℓ_2^m-ball of radius $\sqrt{\widehat{\mu}_m}$. It is known [22] that there exists a 1/2 packing of the ℓ_2^m-ball which has at least $M = 4^m$ elements, all of which have unit norm. By rescaling this packing by $\frac{\sigma_0}{\sqrt{n}}$, we obtain a collection of M vectors $\{\theta^1, \ldots, \theta^M\}$ such that

$$\|\theta^i\|_2^2 = \frac{\sigma_0^2}{n} \quad \text{and} \quad \|\theta^i - \theta^j\|_2^2 \geq \frac{\sigma_0^2}{4n}, \quad \text{for all } i \neq j \in [M].$$

The condition $m \leq (c_0 n)^{\frac{1}{2\alpha}}$ implies that $\|\theta^i\|_2^2 \leq (c_0 \sigma_0^2) m^{-2\alpha} \leq \widehat{\mu}_m$, where the second inequality follows since by assumption (A1) we can take σ_0^2 sufficiently small. Thus, these vectors are also contained within the ellipse $\mathcal{E}$, even after we rescale them further by 1/4, which establishes the claim.

(b) This part makes use of the elementary inequality

$$k \log k - k \overset{(\ell)}{\leq} \sum_{j=1}^{k} \log j \overset{(u)}{\leq} (k+1)\log(k+1) - (k+1). \tag{81}$$

We use known results on the entropy numbers of diagonal operators, in particular for the operator mapping the ℓ_2-ball to the ellipse $\mathcal{E}$. By assumption, we have $\widehat{\mu}_j j^{2\alpha} \in [c_\ell,\, c_u]$ for all $j = 1, 2, \ldots, m$. By Proposition 1.3.2 of [9] with $p = 2$, we have

$$\begin{aligned} \log N_\Phi(\epsilon; \mathcal{E}) &\leq \max_{k=1,2,\ldots,m} \big\{ \frac{1}{2} \sum_{j=1}^{k} \log \widehat{\mu}_j + k \log(1/\epsilon) \big\} + \log 6 \\ &\leq \max_{k=1,2,\ldots,m} \big\{ -\alpha \sum_{j=1}^{k} \log j + k \log(1/\epsilon) \big\} + \log(6 c_u) \\ &\leq \max_{1 \leq k \leq m} f(k) + \log(6 c_u), \end{aligned}$$

where $f(k) = \alpha(k - k\log k) + k \log(1/\epsilon)$. Since $f'(k) = -\alpha \log k + \log(1/\epsilon)$, the optimum is achieved for $k^* = (1/\epsilon)^{1/\alpha}$, and has value $f(k^*) = \alpha(1/\epsilon)^{1/\alpha}$, which establishes the claim.

In the other direction, for all $k \in \{1, 2, \ldots, m\}$, we have

$$\log M_\Phi(\epsilon; \mathcal{E}) \geq \frac{1}{2} \sum_{j=1}^{k} \log \widehat{\mu}_j + k \log(1/\epsilon) \geq -\alpha \sum_{j=1}^{k} \log j + k \log(1/\epsilon) + \log c_\ell.$$

Using the lower bound (81)(u), we obtain

$$\log M_\Phi(\epsilon; \mathcal{E}) \geq \alpha((k+1) - (k+1)\log(k+1)) + k \log(1/\epsilon) + \log c_\ell.$$

The choice $k + 1 = (1/\epsilon)^{1/\alpha}$, which is valid under the given condition $(1/\epsilon)^{1/\alpha} \leq m - 1$, yields the claim.

E.3 Proof of Lemma 16

Any function f in the set Ψ_1^m has the form $f = \sum_{j=1}^m a_j \psi_j$ for a vector of coefficients $a \in \mathbb{R}^m$ such that $\sum_{j=1}^m a_j^2 / \mu_j \leq 1$. If $g = \sum_{j=1}^m b_j \psi_j$ is a second function, then we have $\|f - g\|_{L^2} = \|a - b\|_2$ by construction. Thus, the problem is equivalent to bounding the covering/packing numbers of the m-dimensional ellipse specified by the eigenvalues $\{\mu_1, \ldots, \mu_m\}$. The claim thus follows from the proof of Lemma 15(b).

F Suprema involving Gaussian products

Given a diagonal matrix $Q := \operatorname{diag}(\gamma_1, \ldots, \gamma_m) \in \mathbb{R}^{m \times m}$, this appendix provides bounds on $\|Q^{1/2}\xi\|_2$ where $\xi \in \mathbb{R}^m$ is some random vector (product of Gaussians in particular). The following bound, which follows from Jensen's inequality, is useful:

$$\mathbb{E}\,\|Q^{1/2}\xi\|_2 \le \sqrt{\mathbb{E}\,\|Q^{1/2}\xi\|_2^2} = \sqrt{\operatorname{tr}(Q\Sigma_\xi)}, \quad \text{where } \Sigma_\xi := \mathbb{E}\,\xi\xi^T. \tag{82}$$

We prove a bound for the random vector $\xi := n^{-1}\sum_{i=1}^n \beta_i w_i \in \mathbb{R}^m$, where $\beta_i \sim N(0,1)$, independent of $w_i \sim N(0, I_m)$, and the pairs (β_i, w_i) i.i.d. for $i = 1, \dots, n$.

Lemma 17. *For all $t \ge 0$, we have*

$$\mathbb{P}\bigg[\frac{\|Q^{1/2}\sum_{i=1}^n \beta_i w_i\|_2}{\|\beta\|_2} > \sqrt{\operatorname{tr}(Q)} + t\sqrt{|\!|\!|Q|\!|\!|}\bigg] \le \exp(-t^2/2), \tag{83}$$

where $\beta = (\beta_1, \dots, \beta_n)$.

Proof. Define $\theta := \beta/\|\beta\|_2$, and observe that θ is uniformly distributed on the sphere S^{n-1}, independent of (w_i); we use σ^{n-1} to denote this uniform distribution. The claim is a deviation bound for $\|Q^{1/2}\sum_{i=1}^n \theta_i w_i\|_2$. With θ held fixed, we have $\widetilde{w} := \sum_{i=1}^n \theta_i w_i \sim N(0, I_m)$. The map $\widetilde{w} \mapsto \|Q^{1/2}\widetilde{w}\|_2$ is Lipschitz, from ℓ_2^m to $\mathbb{R}$, with Lipschitz constant bounded by $|\!|\!|Q^{1/2}|\!|\!| = \sqrt{|\!|\!|Q|\!|\!|}$. Hence, by concentration of the canonical Gaussian measure in $\mathbb{R}^m$, with θ held fixed, we have

$$\mathbb{P}\big[\|Q^{1/2}\widetilde{w}\|_2 - \mathbb{E}\,\|Q^{1/2}\widetilde{w}\|_2 \ge t\sqrt{|\!|\!|Q|\!|\!|}\big] \le \exp(-t^2/2).$$

Since this bound holds for all realizations of θ, the tower property implies that the same bound holds unconditionally. Finally, from the bound (82), we have $\mathbb{E}\,\|Q^{1/2}\widetilde{w}\|_2 \le \sqrt{\operatorname{tr}(Q)}$, from which the claim follows. □

We now turn to bounding $\|Q^{1/2}(n^{-1}\sum_i \eta_i w_i - u)\|_2$, where $u \in \mathbb{R}^m$ is some fixed vector. Let us patch u with $u_2, \dots, u_m$ so that $\{u, u_2, \dots, u_m\}$ is an orthonormal basis for ℓ_2^m. Let us define the function $\zeta : \mathbb{R}^n \backslash \{0\} \to \mathbb{R}$ as $\zeta(x) := \frac{n^{-1}\|x\|_2^2 - 1}{n^{-1}\|x\|_2}$. With this notation, we have the following:

Lemma 18. *Let $u \in S^{m-1}$ and assume that $U := (u\ U_2) = (u\ u_2\ \cdots\ u_m) \in \mathbb{R}^{m\times m}$ is orthogonal. Let $(w_i, \eta_i) \in \mathbb{R}^{m+1}$ be i.i.d. Gaussian random vectors for $i = 1, \dots, n$ with distribution*

$$\begin{bmatrix} w_i \\ \eta_i \end{bmatrix} \sim N\bigg(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_m & u \\ u^T & 1 \end{bmatrix}\bigg).$$

Then for all $t \ge 0$,

$$\mathbb{P}\bigg[\big\|Q^{1/2}\,(n^{-1}\sum_{i=1}^n \eta_i w_i - u)\big\|_2 \ge \big(1 + \frac{\|\eta\|_2}{\sqrt{n}}\big)\,\big(\sqrt{\frac{\operatorname{tr}(Q)}{n}} + t\sqrt{|\!|\!|Q|\!|\!|}\big)\bigg] \le 2\exp(-n\frac{t \wedge t^2}{16}),$$

where $\eta = (\eta_1, \dots, \eta_n)$.

Proof. Since the pair (w_i, η) is jointly Gaussian, vectors $\{w_i\}$ conditioned on $\eta = (\eta_i)$ are i.i.d. Gaussian with $\mathbb{E}\,[w_i \,|\, \eta_i] = \eta_i u$ and $\operatorname{cov}(w_i \,|\, \eta_i) = I_m - uu^T$. Consequently, conditioned on η, the variable $\widehat{w}_\eta := n^{-1}\sum_i \eta_i w_i - u$ is Gaussian with mean $u(n^{-1}\|\eta\|_2^2 - 1)$ and covariance $n^{-2}\|\eta\|_2^2(I_m - uu^T)$. Consequently, for $\widetilde{w}_\eta := \widehat{w}_\eta/(n^{-1}\|\eta\|_2)$, we have

$$U^T\widetilde{w}_\eta \sim N\bigg(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\zeta(\eta), \begin{bmatrix} 0 & 0 \\ 0 & I_{m-1} \end{bmatrix}\bigg),$$

where we have used $U^T u = \binom{1}{0}$. Note that $U^T \widetilde{w}_\eta$ is actually a degenerate Gaussian vector, so that we can write $U^T \widetilde{w}_\eta = \big(\zeta(\eta),\ w'\big)$, for some $w' \sim N(0, I_{m-1})$.

Defining $\widetilde{Q} := U^T Q U$, we have

$$\|Q^{1/2}\widetilde{w}_\eta\|_2 = \|U^T Q^{1/2} U U^T \widetilde{w}_\eta\|_2 = \|\widetilde{Q}^{1/2}\, U^T \widetilde{w}_\eta\|_2 = \Big\|\widetilde{Q}^{1/2}\begin{bmatrix}\zeta(\eta)\\ w'\end{bmatrix}\Big\|_2.$$

The map $w' \mapsto \big\|\widetilde{Q}^{1/2}\big[\begin{smallmatrix}\zeta(\eta)\\ w'\end{smallmatrix}\big]\big\|_2$ is Lipschitz, from ℓ_2^{m-1} to $\mathbb{R}$, with Lipschitz constant bounded by $\|\widetilde{Q}^{1/2}\| = \|Q^{1/2}\| = \sqrt{\|Q\|}$. By concentration of canonical Gaussian measure in $\mathbb{R}^{m-1}$, we have

$$\mathbb{P}\big[\|Q^{1/2}\widetilde{w}_\eta\|_2 - \mathbb{E}\,\|Q^{1/2}\widetilde{w}_\eta\|_2 > t\sqrt{\|Q\|}\ \big|\ \eta\big] \le \exp(-t^2/2).$$

Define the function $\kappa(\eta) := \langle\!\langle Q,\ I_m + (\zeta^2(\eta) - 1)uu^T\rangle\!\rangle$. Applying the inequality (82) with $\xi = \big[\begin{smallmatrix}\zeta(\eta)\\ w'\end{smallmatrix}\big]$ and $\widetilde{Q}$ instead of Q, we obtain

$$\mathbb{E}\ \|Q^{1/2}\widetilde{w}_\eta\|_2 = \mathbb{E}\ \Big\|\widetilde{Q}^{1/2}\begin{bmatrix}\zeta(\eta)\\ w'\end{bmatrix}\Big\|_2 \le \Big\{\operatorname{tr}\Big(\widetilde{Q}\begin{bmatrix}\zeta^2(\eta) & 0\\ 0 & I_{m-1}\end{bmatrix}\Big)\Big\}^{1/2} = \sqrt{\kappa(\eta)}.$$

Since $Q \succeq 0$, we have

$$\kappa(\eta) = \operatorname{tr} Q + [\zeta^2(\eta) - 1]u^T Q u \ \le\ \operatorname{tr} Q + \zeta^2(\eta)\, u^T Q u\ \le \operatorname{tr}(Q) + \zeta^2(\eta)\|Q\|.$$

Applying the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ yields $\sqrt{\kappa(\eta)} \le \sqrt{\operatorname{tr}(Q)} + |\zeta(\eta)|\sqrt{\|Q\|}$. Consequently, we have shown the conditional bound,

$$\mathbb{P}\Big\{\frac{\|Q^{1/2}\,(n^{-1}\sum_{i=1}^n \eta_i w_i - u)\|_2}{n^{-1}\|\eta\|_2} > \sqrt{\operatorname{tr}(Q)} + (\sqrt{n}t + |\zeta(\eta)|)\sqrt{\|Q\|}\ \Big|\ \eta\Big\} \le \exp(-nt^2/2). \quad (84)$$

By χ^2-tail bounds, we have $\mathbb{P}\big[|\frac{\|\eta\|_2^2}{n} - 1| \ge t\big] \le \exp(-n\frac{t\wedge t^2}{16})$. Conditioned on the complement of this event, we have $|\zeta(\eta)| \le \frac{t}{n^{-1}\|\eta\|_2}$, and hence conditioning also on the complement of the event in bound (84), we are guaranteed that

$$\begin{aligned}\|Q^{1/2}\,(n^{-1}\sum_{i=1}^n \eta_i w_i - u)\|_2 &\le n^{-1}\|\eta\|_2\Big\{\sqrt{\operatorname{tr}(Q)} + (\sqrt{n}t + \frac{t}{n^{-1}\|\eta\|_2})\sqrt{\|Q\|}\Big\}\\ &\le \big(1 + \frac{\|\eta\|_2}{\sqrt{n}}\big)\ \big(\sqrt{\frac{\operatorname{tr}(Q)}{n}} + t\sqrt{\|Q\|}\big),\end{aligned}$$

with probability at least $1 - 2\exp(-n\frac{t\wedge t^2}{16})$. □

G Bounding an operator norm of a Gaussian matrix

Given a sequence positive numbers $\{\gamma_i\}_{i=1}^m$, consider the $\mathcal{E}_\gamma := \{v \in \mathbb{R}^m : \sum_{i=1}^m \gamma_i^{-1} v_i^2 \le 1\}$. In this appendix, we derive an upper bound on the operator norm of a standard Gaussian random matrix $W \in \mathbb{R}^{n\times m}$, viewed as an operator from $\mathbb{R}^m$ equipped with the norm induced by $\mathcal{E}_\gamma$, to $\mathbb{R}^n$ equipped with the standard Euclidean norm $\|\cdot\|_2$.

Lemma 19. *Let $W \in \mathbb{R}^{n\times m}$ be a standard Gaussian matrix. Then for all $t \ge 0$,*

$$\mathbb{P}\big[\sup_{v\,\in\,\mathcal{E}_\gamma}\|Wv\|_2 > \sqrt{\|\gamma\|_1} + (\sqrt{n} + t)\sqrt{\|\gamma\|_\infty}\big] \le \exp\big(-\frac{t^2}{2}\big). \quad (85)$$

Proof. Let $S^{n-1} := \{u \in \mathbb{R}^n \mid \|u\|_2 = 1\}$ denote the Euclidean unit sphere in $\mathbb{R}^n$. Defining $\mathcal{S} = \{s = (u,v) \mid u \in S^{n-1},\, v \in \mathcal{E}_\gamma\}$, consider the Gaussian process $\{Z_s\}_{s\,\in\,\mathcal{S}}$ where $Z_s = \langle\!\langle W,\; uv^T \rangle\!\rangle$. By construction, we have $\sup_{v\,\in\,\mathcal{E}_\gamma} \|Wv\|_2 = \sup_{s\in\mathcal{S}} Z_s$. Our approach is to use Slepian's comparison for Gaussian processes [21] in order to bound $\mathbb{E}[\sup_{s\in\mathcal{S}} Z_s]$ by $\mathbb{E}[\sup_{s\in\mathcal{S}} X_s]$, where X_s is a second Gaussian process. Concretely, we define $X_s := \sqrt{\|\gamma\|_\infty}\langle u, g\rangle + \langle v, h\rangle$, where g and h are independent canonical Gaussian vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Let $s = (u,v)$ and $s' = (u',v')$ belong to $\mathcal{S}$; by an elementary calculation, we have

$$\mathbb{E}\,[(Z_s - Z_{s'})]^2 = \lvert\lvert\lvert uv^T - u'v'^T \rvert\rvert\rvert_{\mathrm{HS}}^2 \;\le\; \|\gamma\|_\infty \|u - u'\|_2^2 + \|v - v'\|_2^2 = \mathbb{E}\,[(X_s - X_{s'})^2],$$

Consequently, we may apply Slepian's lemma to conclude

$$\begin{aligned}\mathbb{E}\,[\sup_{s\in\mathcal{S}} Z_s] \;\le\; \mathbb{E}\,[\sup_{s\in\mathcal{S}} X_s] \;&=\; \sqrt{\|\gamma\|_\infty}\,\mathbb{E}[\sup_{u\in S^{n-1}} \langle u, g\rangle] + \mathbb{E} \sup_{v\,\in\,\mathcal{E}_\gamma} \langle v, h\rangle \\ &= \sqrt{\|\gamma\|_\infty}\,\big(\mathbb{E}\,\|g\|_2\big) + \mathbb{E}\,\|Q^{1/2}h\|_2 \\ &\le \sqrt{\|\gamma\|_\infty}\,\sqrt{n} + \sqrt{\|\gamma\|_1},\end{aligned}$$

where the final inequality follows by Jensen's inequality, and the relation $\mathrm{tr}(Q) = \|\gamma\|_1$.

Finally, we note that $\|W\|_{\mathcal{E}_\gamma, B_2} = \sup_{v\,\in\,\mathcal{E}_\gamma} \|Wv\|_2$ is a Lipschitz function of the Gaussian matrix W, viewed as a vector in ℓ_2^{mn} with Lipschitz constant $\sqrt{\|\gamma\|_\infty}$. Indeed, it is straighforward to verify that $\sup_{v\,\in\,\mathcal{E}_\gamma} \|Wv\|_2 - \sup_{v'\,\in\,\mathcal{E}_\gamma} \|W'v'\|_2 \;\le\; \lvert\lvert\lvert W - W' \rvert\rvert\rvert_{\mathrm{HS}} \sqrt{\|\gamma\|_\infty}$ so that the claim follows by concentration of the canonical Gaussian measure in ℓ_2^{mn} (e.g., see Ledoux [20]). □

H A uniform law

In this appendix, we state and prove a technical lemma used in parts of our analysis. Consider some subset $\mathcal{D}$ of $\mathbb{R}^m$. Let ν be an index taking values in some index set $\mathcal{I}$. We assume that ν is indexing a collection of random (noise) matrices Δ_ν. Suppose that there is a collection of nonnegative nondecreasing (possibly random) functions $\mathcal{G}_\nu : [0,\infty) \to [0,\infty)$ such that for all $t \ge 0$ and $\nu \in \mathcal{I}$

$$\mathbb{P}\Big\{ \sup_{v\,\in\,\mathcal{D},\; \|v\|_2\,\le\, t} G(v;\Delta_\nu) > \mathcal{G}_\nu(t) \Big\} \le c_1 \exp[-c_2\, \theta_\nu (t \wedge t^2)], \tag{86}$$

where $\theta_\nu, \nu \in \mathcal{I}$ are some positive numbers and G is some function.

Lemma 20. *Under (86) and for any collection* $\{t_\nu\}_{\nu\,\in\,\mathcal{I}}$ *such that* $\inf_{\nu\,\in\,\mathcal{I}} \theta_\nu(t_\nu \wedge t_\nu^2) > 0$*, we have for any* $\nu \in \mathcal{I}$*,*

$$\sup_{v\,\in\,\mathcal{D}} \big[G(v;\Delta_\nu) - \mathcal{G}_\nu(2\|v\|_2)\big] \le \mathcal{G}_\nu(2t_\nu). \tag{87}$$

with probability at least $1 - \tilde{c}_1 \exp[-c_2\, \theta_\nu (t_\nu \wedge t_\nu^2)]$.

Proof. The proof is based on a peeling argument (e.g., [31]). Define $c := \inf_{\nu\,\in\,\mathcal{I}} \theta_\nu t_\nu^2$, and fix some $\nu \in \mathcal{I}$. First, note that as v varies over $\mathcal{D}$, the function $v \mapsto \|v\|_2 \vee t_\nu$ varies over $[t_\nu, \infty)$. Define, for $s \in \{1, 2, \dots\}$,

$$\mathcal{D}_s := \Big\{ v \in \mathcal{D} \,:\, 2^{s-1} t_\nu \le \big(\|v\|_2 \vee t_\nu\big) < 2^s t_\nu \Big\}.$$

We have $\mathcal{D} = \bigcup_{s=1}^{\infty} \mathcal{D}_s$. If there exists $v \in \mathcal{D}$ such that

$$G(v, \Delta_\nu) > \mathcal{G}_\nu\big(2\|v\|_2 \vee 2t_\nu\big), \tag{88}$$

then there exist $s \in \{1, 2, \dots\}$ and $\mathcal{D}_s \ni v$ such that (88) holds for v. Using union bound,

$$\mathbb{P}\Big(\exists v \in \mathcal{D} \,:\, G(v, \Delta_\nu) > \mathcal{G}_\nu\big(2\|v\|_2 \vee 2t_\nu\big)\Big) \;\le\; \sum_{s=1}^{\infty} \mathbb{P}\Big(\exists v \in \mathcal{D}_s \,:\, G(v, \Delta_\nu) > \mathcal{G}_\nu\big(2\|v\|_2 \vee 2t_\nu\big)\Big).$$

For $v \in \mathcal{D}_s$, (88) implies

$$G(v, \Delta_\nu) > \mathcal{G}_\nu\big(2\|v\|_2 \vee 2t_\nu\big) \ge \mathcal{G}_\nu(2\, 2^{s-1} t_\nu) = \mathcal{G}_\nu(\, 2^s t_\nu)$$

where we have used $\mathcal{G}_\nu$ being increasing. Since $\mathcal{D}_s \subset \{v : \|v\|_2 < 2^s t_\nu\}$, we conclude that

$$\begin{aligned}
\mathbb{P}\Big(\exists v \in \mathcal{D} \,:\, G(v, \Delta_\nu) > \mathcal{G}_\nu\big(2\|v\|_2 \vee 2t_\nu\big)\Big) \;&\le\; \sum_{s=1}^{\infty} \mathbb{P}\Big(\sup_{\substack{v \in \mathcal{D}, \\ \|v\|_2 \,<\, 2^s t_\nu}} G(v, \Delta_\nu) > \mathcal{G}_\nu(2^s t_\nu)\Big) \\
&\le\; \sum_{s=1}^{\infty} \exp[-\theta_\nu\, 2^s (t_\nu \wedge\, t_\nu^2)]
\end{aligned}$$

from assumption (86). The last summation is bounded above by

$$\sum_{k=1}^{\infty} \exp[-\theta_\nu\, k\, (t_\nu \wedge t_\nu^2)] = \frac{e^{-\theta_\nu (t_\nu \wedge t_\nu^2)}}{1 - e^{-\theta_\nu (t_\nu \wedge t_\nu^2)}} \le \frac{e^{-\theta_\nu (t_\nu \wedge t_\nu^2)}}{1 - e^{-c}} = C\, e^{-\theta_\nu (t_\nu \wedge t_\nu^2)}.$$

We get the assertion by noting that for $a, b \ge 0$, $\mathcal{G}_\nu(a \vee b) = \mathcal{G}_\nu(a) \vee \mathcal{G}_\nu(b) \le \mathcal{G}_\nu(a) + \mathcal{G}_\nu(b)$ because $\mathcal{G}_\nu$ is assumed to be nondecreasing and nonnegative. □

I Some useful matrix-theoretic inequalities

Fan's inequality states that for symmetric matrices A and B and eigenvalues ordered as $\lambda_1(A) \ge \ldots \ge \lambda_m(A)$ (and similarly for B), we have $\operatorname{tr}(AB) \le \sum_{i=1}^{m} \lambda_i(A)\lambda_i(B)$. As a consequence, for a symmetric matrix B and symmetric matrix $A \succeq 0$, we have

$$\lambda_{\min}(B) \operatorname{tr}(A) \;\le\; \operatorname{tr}(AB) \quad\le\; \lambda_{\max}(B) \operatorname{tr}(A). \tag{89}$$

It follows that for a symmetric matrix $D \succeq 0$ and $R \in \mathbb{R}^{r \times r}$, we have

$$\lambda_{\min}(R^T R) \operatorname{tr}(D) \;\le\; \operatorname{tr}(DRR^T) \;=\; \operatorname{tr}(R^T D R) \quad\le\; \lambda_{\max}(R^T R) \operatorname{tr}(D), \tag{90}$$

where we have used the fact that $R^T R$ and RR^T have the same eigenvalues.

For $B \succeq 0$, we have

$$\lambda_{\min}(B)\, \lambda_j(R^T R) \;\le\; \lambda_j(R^T B R) \;\le\; \lambda_{\max}(B)\, \lambda_j(R^T R), \tag{91}$$

which can be established using the classical min-max formulation of the j^{th} eigenvalue—namely

$$\lambda_j(C) = \max_{\mathcal{M}\colon \dim(\mathcal{M}) = j} \;\; \min_{x \,\in\, \mathcal{M} \,\cap\, S^{r-1}} z^T C z \tag{92}$$

where the maximum is taken over all j-dimensional subspaces of $\mathbb{R}^k$. Finally, the inequality (91) implies that

$$\|R^T B R\|_{HS} \;\le\; \|B\|_2\, \|R^T R\|_{HS}. \tag{93}$$

References

[1] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, 37(5B):2877–2921, October 2009.

[2] A. A. Amini and M. J. Wainwright. Approximation properties of certain operator-induced norms on Hilbert spaces. Technical report, UC Berkeley, 2010.

[3] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, Norwell, MA, 2004.

[4] P. Besse and J. O. Ramsay. Principal components analysis of sampled functions. *Psychometrika*, 51(2):285–311, June 1986.

[5] R. Bhatia. *Matrix Analysis*. Springer, 1996.

[6] G. Boente and R. Fraiman. Kernel-based functional principal components. *Statistics & Probability Letters*, 48(4):335–345, 2000.

[7] D. Bosq. *Linear Processes in Function Spaces: Theory and Applications*. Springer, 2000.

[8] H. Cardot. Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonparametric Statist.*, 12(4):503–538, 2000.

[9] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, 1990.

[10] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivariate Anal.*, 12(1):136–154, March 1982.

[11] K. R. Davidson and S. J. Szarek. Handbook of the geometry of banach spaces. In W B Johnson and J Lindenstrauss, editors, *Chapter 8 Local operator theory, random matrices and Banach spaces*, volume 1, pages 317–366. Elsevier Science B.V., 2001.

[12] P. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, USA, 2002.

[13] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.

[14] C. Gu. *Smoothing spline ANOVA models*. Springer Series in Statistics. Springer, New York, NY, 2002.

[15] P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society*, 68(1):109–126, February 2006.

[16] P. Hall, H.-G. Müller, and J.-L. Wang. Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.*, 34(3):1493–1517, June 2006.

[17] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, April 2001.

[18] I. M. Johnstone and A. Lu. Sparse principal components, 2004.

[19] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[20] M. Ledoux. *The Concentration of Measure Phenomenon.* American Mathematical Society, 2001.

[21] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer-Verlag, New York, NY, 1991.

[22] J. Matousek. *Lectures on discrete geometry.* Springer-Verlag, New York, 2002.

[23] S. Mendelson. Geometric Parameters of Kernel Machines. *Lecture Notes In Computer Science*, 2375:29—-43, 2002.

[24] D. Paul and I. Johnstone. Augmented sparse principal component analysis for high-dimensional data, 2008.

[25] S. Pezzulli and B.W. Silverman. Some properties of smoothed principal components analysis for functional data. *Comput Stat.*, 8(1):1–16, 1993.

[26] A. Pinkus. *N-Widths in Approximation Theory.* Springer, New York, 1985.

[27] J. Ramsay and B. W. Silverman. *Functional Data Analysis.* Springer, 2005.

[28] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis.* Springer, 2002.

[29] J. A. Rice and B. W. Silverman. Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves. *Journal of the Royal Statistical Society*, 53(1):233 – 243, 1991.

[30] B. W. Silverman. Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, 24(1):1–24, February 1996.

[31] S. A. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2009.

[32] G. Wahba. *Spline models for observational data.* CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PN, 1990.

[33] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[34] F. Yao, H.-G. Müller, and J.-L. Wang. Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, 100(470):577–590, June 2005.

[35] B. Yu. Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, Berlin, 1997.